Information Systems 40 (2014) 11–31

Contents lists available at ScienceDirect

Information Systems

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

Using a knowledge learning framework to predict errors in database design

Sabah Currim [a,*], Sudha Ram [b], Alexandra Durcikova [c], Faiz Currim [b]

[a] *Enterprise Information & Analytics, University of Arizona, USA*
[b] *Department of Management Information Systems, University of Arizona, USA*
[c] *Division of Management Information Systems, University of Oklahoma, USA*

ARTICLE INFO

Article history:
Received 8 March 2012
Received in revised form
18 June 2013
Accepted 6 August 2013
Available online 15 August 2013

Keywords:
Database design
Modeling expertise
Entity Relationship modeling
Revised Bloom's taxonomy
Analysis of errors

ABSTRACT

Conceptual data modeling is a critical but difficult part of database development. Little research has attempted to find the underlying causes of the cognitive challenges or errors made during this stage. This paper describes a Modeling Expertise Framework (MEF) that uses modeler expertise to predict errors based on the revised Bloom's taxonomy (RBT). The utility of RBT is in providing a classification of cognitive processes that can be applied to knowledge activities such as conceptual modeling. We employ the MEF to map conceptual modeling tasks to different levels of cognitive complexity and classify current modeler expertise levels. An experimental exercise confirms our predictions of errors. Our work provides an understanding into why novices can handle entity classes and identifying binary relationships with some ease, but find other components like ternary relationships difficult. We discuss implications for data modeling training at a novice and intermediate level, which can be extended to other areas of Information Systems education and training.

1. Introduction

Researchers and practitioners continue to be concerned with reducing design errors, including those in database design [1–3], owing to the high cost of faulty software which is estimated to be in the range of tens of billions of dollars per year [4]. Forty-five to sixty-five percent of all errors are made during design [5]. The cost of fixing an error in development is proportional to the time the error remains in the process; the expected savings increase between one and two orders of magnitude when an error is fixed near the creation point, rather than at the implementation stage [6,7]. It is therefore essential to remove errors early in the design phase, e.g., during conceptual design. Before we can eliminate errors, it is constructive to identify and analyze them so we have an understanding of what errors/mistakes designers make and why. This paper examines modeler expertise and its impact on errors made in conceptual database design.

Owing to the significant cost of errors, previous research has looked at the impact of a variety of factors that cause design errors including challenges in requirements gathering, choice of modeling grammar, the application domain, and designers' background [8]. Arguably, some types of errors will be harder to eliminate, for example, those caused by changing user requirements. Our attention is on errors that can be attributed to lack of modeling expertise. By expertise we mean the level of knowledge and skills achievement in conceptual modeling formalism (in our case, the Entity Relationship or ER Model) required to effectively create a schema.

Examining the impact of modeling expertise is important for a variety of reasons. First, by establishing a predictive relationship between levels of modeling expertise and errors, steps can be taken to improve the quality of

* Correspondence to: Enterprise Information & Analytics, 1077 N. Highland Avenue, University of Arizona, Tucson AZ 85721, USA. Tel.: +1 520 626 1176.
E-mail address: scurrim@email.arizona.edu (S. Currim).

0306-4379/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.is.2013.08.001

generated schema. Second, it is necessary to understand the cause of an error type to facilitate action to prevent its recurrence [1], with the ultimate goal of reducing the total number of errors. Third, unlike some factors (e.g., modeling techniques), modeling expertise has not been sufficiently examined in existing literature as an antecedent to errors [8]. Finally, since expertise is a factor that can be influenced through training [9], when it is combined with a granular breakdown of error classes, research can empirically establish the nature of training required for different levels of expertise.

Past research shows that failure to understand the impact of modeling expertise on errors leads to incomplete and incorrect application of data modeling [10]. This motivates research to go beyond describing errors and find ways to predict them and reduce their recurrence. One example of incorrect and incomplete application of data modeling is the ER-stencil in Microsoft Visio which confounds ER (semantic) and relational (logical) design principles. The stencil does not support many-to-many (M-M) binary, ternary, or higher degree interaction relationships, and it includes foreign identifiers within a class to represent relationships between entity classes. While practitioners may see this ER-relational combination as a convenience, novices may incorrectly assume that M-M binary and ternary relationships are erroneous constructs, and may not grasp that foreign identifiers are redundant since a relationship already captures the association among participating entity classes. ER modeling is not well understood and, as this CASE tool example illustrates, can be used incorrectly in practice. Nevertheless, database development employing ER/EER modeling has been shown to result in better performance than starting directly with relational design (summarized in Topi and Ramesh [8]).

Our paper proposes a knowledge framework that links modeling expertise to errors in ER modeling. In this study, we restrict our scope to knowledge-based errors in conceptual (semantic) data modeling,[1] an important part of database design. Specifically, the objective of our research is to better understand *why novices make modeling expertise-related errors in ER modeling* (this is not a statement as to the efficacy of ER versus UML; designing the study required us to make a choice, but the same methodology can be used for UML). To address our objective, the revised Bloom's taxonomy (RBT) knowledge framework was applied to ER modeling to develop our Modeling Expertise Framework (MEF). The MEF allowed us to predict types of errors in conceptual modeling when the expertise level was known. An experiment designed to evaluate the framework confirmed its validity. We also develop an algorithm to validate a conceptual test schema, given a solution schema that builds on the fine grained classification of errors (see Table 2) to identify errors and analyze the violations corresponding to each RBT expertise level. In the interest of space, we present our work on interaction relationships in this paper. The complete set of ER constructs is discussed in the dissertation work of one of the authors.

[1] We refer to *semantic data modeling* instead of *conceptual data modeling* to minimize overloading the term "conceptual"; the term is also used in the underlying theory framework adopted by this paper.

2. Theoretical background and hypotheses

2.1. Errors in data modeling

Batra et al. [11] identified three major factors that influence database design performance for ER and relational design: data model choice, modeling task, and modeler characteristics. A detailed discussion of these factors follows.

A number of studies have examined the impact of the data model on accuracy and completeness of schema created by comparing different grammars, such as the ER model (and its variants), the object-oriented model, and the relational model. These are summarized in past research surveys [8,12]. As new conceptual data modeling grammars were developed, the studies were useful to compare expressiveness of competing techniques and provide guidelines for adoption. Over time, the data management field matured and modeling standards were established, e.g., ER and UML as conceptual grammars, and the Relational Model for logical design. Newer research has moved towards tackling questions of how task and designer characteristics affect modeling performance [8] and how to improve usability [6]. We are interested in the nature of the task as well as the modeler characteristics. Our approach can be used along with Norman's Theory of Action [6] to improve usability at different levels of expertise.

Modeling tasks can be measured along two major dimensions: type of task (e.g., interpreting, validating, and creating diagrams) and task complexity (e.g., structure, difficulty, and time). Several studies have examined a single task type [13], while others worked with multiple task types [2,14] and complexities [15]. In general, the more complex the task is, the worse the performance. Intuitively this makes sense, as more complex modeling tasks require a higher level of expertise to be successfully completed. Studies have typically stayed away from a combination of multiple task types and complexities because they are harder to measure.

Finally, modeler characteristics explored previously include cognitive style [16], application domain knowledge [17], and different aspects of experience. Aspects of experience have been examined including IS [18], data modeling [19], and general modeling and programming [20]. Modeling expertise has not been studied previously, though *experience* may be considered by some researchers as a surrogate for expertise [18,19]. Experience and expertise are related but distinct concepts [21]. For example, a programmer may have multiple years of experience in developing databases, but he may not necessarily be an expert. Likewise, two developers with equal years of experience may have differing levels of expertise in the subject matter. The differences in expertise levels for novices, who have less than a year experience, can vary a fair amount. Topi et al. [8] call for more research in this area because semantic modeling techniques like ER can be challenging to master, and we need to better understand the causes of this difficulty. While previous research has highlighted that novices and experts are at different levels in the same process model [19], we need a classification of the different levels of cognitive complexity [6,12] within ER modeling to understand the reasons behind novice errors.