2000

A Meta-Analysis on Relationship Modeling Accuracy: Comparing Relational and Semantic Models

Qing Cao
University of Nebraska at Lincoln, qcao@unlserve.unl.edu

Fiona Nah
University of Nebraska at Lincoln, fnah2@unl.edu

Keng Siau
University of Nebraska at Lincoln, siauk@mst.edu

A Meta-Analysis on Relationship Modeling Accuracy: Comparing Relational and Semantic Models

Qing Cao, Fiona Nah, Keng Siau
Management Department
University of Nebraska – Lincoln
qcao@unlserve.unl.edu, fnah2@unl.edu, ksiau@unl.edu

Abstract

Semantic data modeling, such as entity-relationship (ER) modeling and extended/enhanced entity-relationship (EER) modeling, has emerged as an alternative to relational data modeling. The majority of research in data modeling suggests that the use of semantic data models leads to better performance. However the findings are not conclusive and sometimes inconsistent. In this research, we investigate modeling relationship correctness in relational and semantic models. The meta-analysis carried out in this research is an attempt to alleviate inconsistent results in previous studies.

Introduction

Modeling is the *sine qua non* of software development (Siau, 1999; Selic, 1999; Booch *et al.* 1999; Rumbaugh *et al.* 1999). As the success of software development projects depends heavily on the accuracy of the data models (Ramakrishnan, 1998; Siau *et al.* 1997), it is important to compare user performance across different data modeling methods.

Numerous data modeling approaches have been introduced in the last three decades. Two of the most popular approaches are the relational model and the semantic models. Relational model is the most popular technique for managing large collections of data (Watson 1999). A relational model consists of a set of relations, where a relation is a mathematical term for a table. Semantic data models emerged in the mid 1970s as an alternative to relational modeling. The main advantage of a semantic data model is that it captures more semantic information than the relational model. The most representative semantic data model is the entity-relationship (ER) model. Often, database designers begin by developing a schema using the ER model and then translate the ER schema to the relational model for implementation (Ullman & Widom 1997). The basic ER model was subsequently extended to include more advanced concepts such as generalization/specialization and aggregation. These improved ER models are known as Extended or Enhanced ER (EER) models.

Since the introduction of semantic models in the mid 1970s, several empirical studies on user performance have examined and compared the relational model and semantic data models (Chan & Lim, 1998; Chan *et al.*, 1998; Siau *et al.*, 1995; Kim & March, 1995; Chan *et al.*, 1993; Batra *et al*., 1990). The majority of research in data modeling suggests that the use of semantic data models leads to better performance, however, the findings on relationship modeling are inconclusive (Batra *et al*., 1990; Jarvenpaa & Machesky, 1989).

Literature Review

Data modeling is widely studied in the MIS area. Empirical studies, especially experimental studies, are commonly used in comparing semantic data modeling and relational data modeling. Leitheiser (1988) studied end-user model comprehension. He found that a semantic model (LDS) was easier to learn and resulted in higher understanding and recall of a database schema than a tabular representation. Batra *et al.* (1990) studied end-user model building using both the EER model and the relational model. They found that the EER model scored higher in modeling binary relationships and one-many-many ternary relationship than the relational model. Siau *et al.* (1995) studied the accuracy of formulating queries for the relational and ER model. They found that subjects using the ER model performed better.

Although the general finding is that the ER model outperforms the relational model, some studies did not find semantic models to be superior to the relational model. For example, in modeling unary relationships, Batra *et al*. (1990) found no significant difference between the two modeling approaches.

A meta-analysis on modeling relationship is carried out in this research to investigate the inconsistencies.

Research Framework

The research model for this study consists of four factors: Human, Data Model, Task, and Performance. The human and task factors are controlled in our research. We only consider Novice subjects for the Human factor and Modeling task for the Task factor. The same research model has been used in numerous prior studies (e.g., Bock & Ryan, 1993; Hardgrave and Dalal, 1995).

Independent Variable

The independent variable, Data Model, has two levels. The two levels are relational model and semantic data models (ER and EER).

Dependent Variable

The dependent variable is Performance. In this study, performance is operationalized as modeling correctness. Batra *et al*. (1990, p.130) defined modeling correctness as "the degree to which a conceptual model approaches the correct solution(s), where the correct solution(s) convey the same semantics about data as the natural language description of the applications".

Several types of relationships are possible. In this research, we consider unary, binary, and ternary relationships. Binary relationships can be further divided into Binary 1:M and Binary M:M relationships. As for ternary relationships, we separate them into Ternary 1:M:M and Ternary M:M:M relationships. Although Ternary 1:1:M relationship is also possible, there is only one study that involved this relationship. As such, Ternary 1:1:M relationship is not included in this research.

Theory and Hypotheses

One of the reference disciplines for data modeling research is Human-Computer Interaction. The Human-Computer-Interface (HCI) model proposed by Booth (1992) is used as the theoretical foundation for this research. In the HCI model, a gulf of execution separates the physical system from the user's goals. According to the model, in order to improve the performance of the users, the gulf needs to be narrowed as much as possible.

In the context of database design, semantic gulf of execution is the distance between the meaning of the conceptual model and the user's knowledge of the real world (Batra *et al*., 1990). While the ER/EER models use special graphical notations to represent the degree and connectivity of relationships, the relational model captures relationships in a more complicated manner.

Since semantic models (ER/EER) capture the characteristics of relationships between entities in a more direct fashion, it is logical to assume that semantic models would better bridge the semantic gulf of execution than the relational model. Hence, it is hypothesized that:

Users of semantic data models (ER/EER), as compared to users of relational model, will perform significantly better in modeling:

H1) unary relationships
H2) binary one-many relationships
H3) binary many-many relationships
H4) ternary one-many-many relationships
H5) ternary many-many-many relationships

Research Methodology

Meta-analysis involves the statistical integration of the results of independent studies and it represents a more rigorous and more precise approach in the summarization and integration of research literatures. The benefit of meta-analysis is two-fold. First, meta-analysis provides the opportunity to remove sampling errors in the studies and hence results in a more accurate summary of the literature. Second, the quantitative measures derived from a meta-analysis include the significance level, the magnitude of the phenomena under study, and the variability of the outcomes.

Meta-analysis Procedure

After a thorough search of databases including ABI/Inform, INFORTAC, and ORMS-WEB, we found 9 previous studies on data modeling that involved the comparison of semantic (ER/ERR) and relational models.

Since the statistical procedures used in the independent studies may vary, metrics for the significance level and the effect size of an independent variable on the dependent variable in the meta-analysis must be established. In this study, p value is used to present the significance level since it appears most often in the data modeling studies. Two commonly used effect size measures in meta-analysis are Cohen's d and Pearson's r (Hwang, 1996). Pearson's r was chosen in this study not only because it is the most commonly used effect size measure in data modeling research but also because its use does not require the demonstration of causal connections. In this research, Pearson's r was either obtained from the findings of the studies or derived from other statistical tests such as F, t, or k^2.

All findings were then grouped based on the hypotheses. The mean effect size and standard deviation were calculated using the formulas proposed by Hunter *et al*. (1990):

Mean effect size (R) = $\frac{\sum N_i r_i}{\sum N_i}$

Sample variance (Var(R)) = $\frac{\sum N_i (r_i - R)^2}{\sum N_i}$

Standard deviation (SD) = $\frac{Var(R)}{M}$

Where N_i = the sample size of independent study i
r_i = the effect size of independent study i
M = the number of studies included in the meta-analysis

A rule of thumb for judging the magnitude of effect sizes was suggested by Cohen (1977) – an effect size of 0.2 represents a small effect, an effect size of 0.5 constitutes a medium effect, and an effect size of 0.8 means a large effect.

Results and Discussion

The test result is displayed in Table 1. For each hypothesis, Table 1 shows the number of studies used in the meta-analysis, cumulative sample size, average effect of the independent studies, mean effect of the meta-analysis, standard deviation of the meta-analysis, and significance level (p-value).

Hypothesis 1 (modeling correctness in unary relationships) is statistically significant at the 0.1 level ($p=0.0929<0.1$), which indicates the superiority of semantic models (ER/ERR) when compared to the relational model in modeling unary relationship. The result is different from the study by (Batra *et al*, 1990) where they found no significant difference between the relational and ER model in terms of modeling correctness of unary relationships. One possible explanation for the inconsistent results is that they had a relatively small sample size (19 subjects) compared to our meta-analysis (423 subjects). In this case, the effect size is increased by 12%.

Hypotheses 2, 3 and 4 (modeling correctness in binary one-many relationships, binary many-many relationships and ternary one-many-many relationships) are supported. This is in line with the numerous prior findings that semantic models outperform the relational model in modeling binary and ternary one-many-many relationships. The findings indicate that the effect size increased by 15%, 16%, and 12% respectively.

For hypothesis 5 (modeling correctness in ternary many-many-many relationships), this study did not produce a statistically significant result -- even though the effect size is increased by 13% (higher magnitude of treatment and higher statistical power). One possible explanation is that the subjects involved in this meta-analysis were novices in data modeling. It is likely that ternary many-many-many relationships were too difficult for them to apprehend.

Conclusion

Research in data modeling has made significant contribution to our understanding of the pros and cons of different data modeling techniques. This meta-analysis study has not only alleviated inconsistencies in relationship modeling correctness but also produced larger effect size across the board, which translates into higher statistical power.

The use of the meta-analysis approach is not without criticism. Firstly and foremost is the "apples and oranges" criticism – the mixing of data that come from studies in diverse settings. To alleviate this problem, we select studies that are in line with our research question. However, some degree of heterogeneity is inevitable -- irrespective of whether meta-analysis or narrative reviews are carried out.

Results / Hypothesis	Number of studies	Sample size	Average effect	Mean effect (Ri)	Standard deviation	p value
Unary relationship	7	423	0.263	0.298	0.032	0.0929
Binary 1:M relationship	9	673	0.713	0.815	0.013	0.0001
Binary M:M relationship	9	613	0.672	0.736	0.009	0.0000
Ternary 1:M:M relationship	9	644	0.532	0.594	0.008	0.0000
Ternary M:M:M relationship	8	550	0.491	0.562	0.011	0.1211

Table 1 – Effect size and significance levels of the results

References

Batra, D., Hoffer, J.A., and Bostrom, R.P. "A comparison of user performance between the relational and the extended entity relationship models in the discovery phase of database design," *Communication of the ACM* (33:2), 1990, pp. 126-139.

Bock, D.B. and Ryan, T. "Accuracy in Modeling with Extended Entity Relationship and Object-oriented Data Models," *Journal of Database Management* (4:4), 1993, pp. 30-39.

Booch, G., Rumbaugh, J., and Jacobson, I. The Unified Modeling Language User Guide, Addison Wesley, 1999.

Booth, P., An Introduction To Human-Computer Interaction, Lawrence Erlbaum Associates, 1992.

Chan, H., Wei, K., and Siau, K. "User-Database Interface: The Effect of Abstraction Levels on Query Performance," *Management Information Systems Quarterly* (17:4), December 1993, pp. 441-464.

Chan, H. and Lim, L. "Database Interfaces: A Conceptual Framework and a Meta-Analysis on Natural Language Studies," *Journal of Database Management* (9:3), 1998, pp. 25-32.

Chan, H., Siau, K., and Wei, K. "The Effect of Data Model, System and Task Characteristics on User Query Performance – An Empirical Study", *Data Base* (29:1), 1998, pp. 31-49.

Cohen, J. Statistical Power Analysis for the Behavioral Sciences, Academic Press, New York, NY, 1977.

Hardgrave, B.C. and Dalal, N.P. "Comparing Object Oriented and Extended Entity Relational Data Model," *Journal of Database Management* (6:3), 1995, pp. 15-21.

Hunter, J.E., Schmidt, F.L., and Jackson, G.B. Meta-Analysis: Cumulating Research Finding Across Studies, Sage Publication, Beverly Hills, CA, 1990.

Hwang, M.I. "The use of meta-analysis in MIS research: promises and problems", *Data Base* (27:3), 1996, pp. 35-48.

Jarvenpaa, S. and Machesky, J. " Data Analysis and Learning: An Experimental Study of Data Modeling Tools," *International Journal of Human Computer Studies*, (31), 1989, pp.367-391.

Kim, Y. and March, S.T. "Comparing Data Modeling Formalisms," *Communication of the ACM* (38:6), 1995, pp. 123-113.

Leitheiser, R. An examination of the effects of alternative schema descriptions on the understanding of database structure and the use of a query language. Ph.D. dissertation, Univ. of Minnesota, Minneapolis, 1988.

Ramakrishnan, R. Database Management Systems. WCB/McGraw-Hill, 1998.

Rumbaugh, J., Jacobson, I., and Booch, G. The Unified Modeling Language Reference Model, Addison Wesley, 1999.

Selic, B. "Turning Clockwise: Using UML in the Real-Time Domain," *Communications of the ACM* (42:10), 1999, pp. 46-54.

Siau, K. "Information Modeling and Method Engineering: A Psychological Perspective," *Journal of Database Management* (10:4), 1999, pp. 44-50.

Siau, K., Chan, H., and Wei, K. "End Users Communication With Systems: The Effects of Conceptual and Logical Interfaces on Visual Query Performance of End Users," *Informatie* (37), November 1995, pp. 660-668.

Siau, K., Chan, H., and Wei, K. "The Effects of Conceptual and Logical Interfaces on Visual Query Performance of End-Users," *Proceedings of the 16th Annual International Conference on Information Systems*, December 1995, Amsterdam, Netherlands, pp. 225-235.

Siau, K., Wand, Y., and Benbasat, I. "A Psychological Study on the Use of Relationship Concept -- Some Preliminary Findings," *Lecture Notes in Computer Science -- Advanced Information Systems Engineering* (932), 1995, pp. 341-354.

Siau, K., Wand, Y., and Benbasat, I. "When Parents Need Not Have Children -- Cognitive Biases in Information Modeling," *Lecture Notes in Computer Science -- Advanced Information Systems Engineering* (1080), 1996, pp. 402-420.

Siau, K., Wand, Y., and Benbasat, I. "The Relative Importance of Structural Constraints and Surface Semantics in Information Modeling," *Information Systems* (22:2&3), 1997, pp. 155-170.

Ullman, J.D. and Widom, J. A First Course in Database Systems, Prentice Hall, 1997.

Watson, R. Data Management: Databases and Organizations, Second Edition, John Wiley & Sons, 1999.