View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by AIS Electronic Library (AISeL)

Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2001 Proceedings

Americas Conference on Information Systems (AMCIS)

December 2001

Dimensional Data Models versus Entity Relationship Models: Does it Make a Difference to End-Users?

Karen Dowling
Arizona State University

David Schuff
Temple University

Robert St. Louis
Arizona State University

Follow this and additional works at: http://aisel.aisnet.org/amcis2001

Recommended Citation
Dowling, Karen; Schuff, David; and St. Louis, Robert, "Dimensional Data Models versus Entity Relationship Models: Does it Make a Difference to End-Users?" (2001). *AMCIS 2001 Proceedings*. 80.
http://aisel.aisnet.org/amcis2001/80

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in AMCIS 2001 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

DIMENSIONAL DATA MODELS VERSUS ENTITY RELATIONSHIP MODELS: DOES IT MAKE A DIFFERENCE TO END-USERS?

Karen Dowling
Arizona State University
karen.dowling@asu.edu

David Schuff
Temple University
schuff@temple.edu

Robert St. Louis
Arizona State University
st.louis@asu.edu

Abstract

The more closely structures approximate the way people think, the easier they are for people to understand, remember, and use. This paper explores whether a dimensional data model is easier to remember than an entity-relationship model of similar size and complexity. A laboratory experiment is conducted to determine which modeling method results in more accurate recall. The results show that users are able to recall the elements of a dimensional data model much more accurately than they are able to recall the elements of an entity-relationship model. Further research is needed to determine whether easier recall translates into easier use.

Keywords: Dimensional data model; star-join schema; entity-relationship diagram, normalized relational schema

Introduction

The old paradigm for information technology (IT) departments was that (1) users request data, (2) IT departments decide which requests to honor, and then (3) some users ultimately get their requested data. Three major problems with this paradigm were that the IT department became a horrific bottleneck, many users were unable to get the data they needed to make decisions, and almost all users got the data they needed too late or in the wrong format. The advent of the data warehouse has changed this paradigm. The resulting paradigm is that (1) the IT department maintains an enterprise data warehouse, and ~~the~~(2) users directly access the data warehouse to obtain the needed data themselves. However, there are problems with this new paradigm. Specifically, the users must be aware of what data is available, and the user interface must be simple enough to enable non-technical end-users to access the data warehouse and retrieve the information themselves.

As a result of this paradigm shift, expenditures on data warehouses have become a significant part of the total IT budget, and are expected to become even more important in the future. More specifically, expenditures on data warehouses are expected to increase from $37.4 billion dollars in 1999 to $148.5 billion dollars in 2003 [Business Intelligence/Data Warehousing Research Program 2000, ch. 6]. This is an increase of over 40 percent per year. The payback from these expenditures is directly related to the effectiveness of the end-user delivery databases. If end-users are unable to ascertain what information is available in the data warehouse then the data warehouse is not effective.

Some authors have argued that normalized relational schemas and SQL are an adequate end-user delivery system. Roland (1998, p. 12), for example, states that "Relational systems provide a development environment that is significantly easier to use than that provided by the previous approaches. The data structures are simple to build and easy to understand and the writing of programs to manipulate them relatively straightforward." Other authors have argued that relational schemas and SQL are not an adequate end-user delivery system. Kimball (1997), for example, states that “dimensional modeling is the only viable technique for designing end-user delivery databases.” Although the pervasiveness of dimensional data marts tends to support Kimball’s statement, there is no research to support the proposition that dimensional models are more easily understood than entity-relationship models. This paper presents the results of an experiment that tests whether dimensional data models are more easily understood than entity-relationship models. This is extremely important because the understandability of the end-user delivery system is a major determinant of the effectiveness of the data warehouse.

Background

The primary difference between dimensional data models (DDMs) and entity-relationship diagrams (ERDs) is the pattern. A DDM organizes the relationships into a pattern around the central fact table. This pattern is called a star schema. An ERD doesn't have this elemental pattern. Research in cognitive science tells us that "organization is a necessary condition for memory" (Mandler, 1967, p. 328). "The structure of human memory appears to be the primary determinant of how individuals encode and elaborate their understanding of the real world" (Weber, 1996, p. 140). There is considerable evidence that humans are able to remember and process only a limited amount of information. Miller (1956) demonstrated that this limitation is the now famous "seven-plus-or-minus-two." Yet humans clearly can process much more than seven pieces of information. To explain this apparent contradiction, Miller suggested that each individual unit of information may contain many subunits. This hierarchical storage of units led to the concept of memory "chunking."

Semantic network theory says that being able to retrieve information requires the retrieval of not only the concepts, but also the links between the concepts (Anderson, 1990). Retrieval and comprehension are closely related. It is much easier to remember (retrieve) information that makes sense than information that does not make sense (e.g., words are easier to remember than random collections of letters). Data models are a form of semantic network. Entities are analogous to concepts. They are sets of attributes representing some perceived concept. Relationships are analogous to links. They represent the associations among entities or the ways the entities relate to one another. The ability to use a data model correctly and to its fullest advantage is predicated on understanding the relationships that exist within that data model. Both ERDs and star schemas are semantic nets. Because of the organization of star schemas, semantic network theory suggests that they will be easier to recall (and therefore be more understandable) than ERDs.

The specific null hypotheses that are tested are:

- *Hypothesis 1: subjects will remember no more relationships in a star schema than in an ERD*
- *Hypothesis 2: subjects will remember no more entities in a star schema than in an ERD*

Semantic network theory tells us that finding the paths between concepts is critical to the ability to comprehend and retrieve concepts, and that not all paths are equally strong in memory – i.e., some may be easier to access. It also tells us that concepts which are closer together are easier to retrieve. Because all entities are a similar distance from the fact table in a star schema but not in an ERD, semantic network theory suggests that both the entities and the relationships should be easier to recall in a star schema. "Distance" in a network refers, not to the length of individual relationships, but to the number of different relationships which must be traversed to join two entities.

The Experiment

To determine how people store memories, cognitive scientists use recall experiments (see Weber, 1996, for a brief review). To determine if subjects would find ERDs or star schemas easier to recall, and therefore, easier to understand, this research conducted two recall experiments. The experiments involved both simple and complex data diagrams. The terms simple and complex diagrams are not absolutes, but are meant to distinguish between the two cases. Semantic network theory states that the larger the network (i.e., the greater the number of nodes and links), the more difficult the memory task. The simple ERD had ten entities and eleven relationships, while the complex ERD had 16 entities and 19 relationships. The simple star schema had ten entities and nine relationships, while the complex star schema had 19 entities and 18 relationships.

It is critical that the ERD and star schema diagrams used in these experiments contain the same information. If one diagram contains more information than the other, then differences in recall could be due to differences in information content rather than differences in information presentation. A set of relevant business questions was constructed, and the paired diagrams were tested for their ability to answer those questions. This resulted in slight differences in the number of entities and relationships in the diagrams. In analyzing the results, the numbers were standardized to allow direct comparison.

Subjects were graduate students enrolled in a database class. The subjects were divided into four groups: simple star schema diagram; simple ERD diagram; complex star schema diagram, complex ERD. They were given a short amount of time to look at the diagram for their group, and then asked to draw it from memory. While there was a time limit to the drawing phase, in fact everyone reproduced as much of the diagram as they could in the allotted time.

The number of correct, missing, and extraneous entities was counted. The number of wrong entities was computed by adding together the number of missing and extraneous entities. Similarly, the number of correct, missing and extraneous relationships

was counted. The number of wrong relationships was computed by adding together the number of missing and extraneous relationships. Because the number of entities and relationships differed in the four diagrams, the counts were standardized (e.g., in the simple ERD there were 10 entities, therefore the number of entities wrong was divided by 10). This allows an easier comparison of the results between treatments.

Results

The results from this experiment were analyzed using NCSS' general linear models procedure. Figure 1 graphically displays the results. From Figure 1 it appears that end-users do not recall entities more accurately from a star schema than from an ERD (the p-value for the diagram-type main effect was .96, and the p-value for the interaction was .44). Hence Hypothesis 2 is not rejected. However, it is very clear from Figure 1 that end-users do recall relationships more accurately from a star schema than from an ERD, and that this difference is most pronounced for complex models (the p-value for the diagram-type main effect was .0000, and the p-value for the interaction was .044). Hence Hypothesis 1 is rejected. For the use of a data model, the relationships are critical. Thus this experiment provides initial justification for Kimball's statement that the star schema is easier to use than the ERD. The issue for future research is whether the ability to more readily understand the star schema translates into an ability to more easily extract information from a DDM.

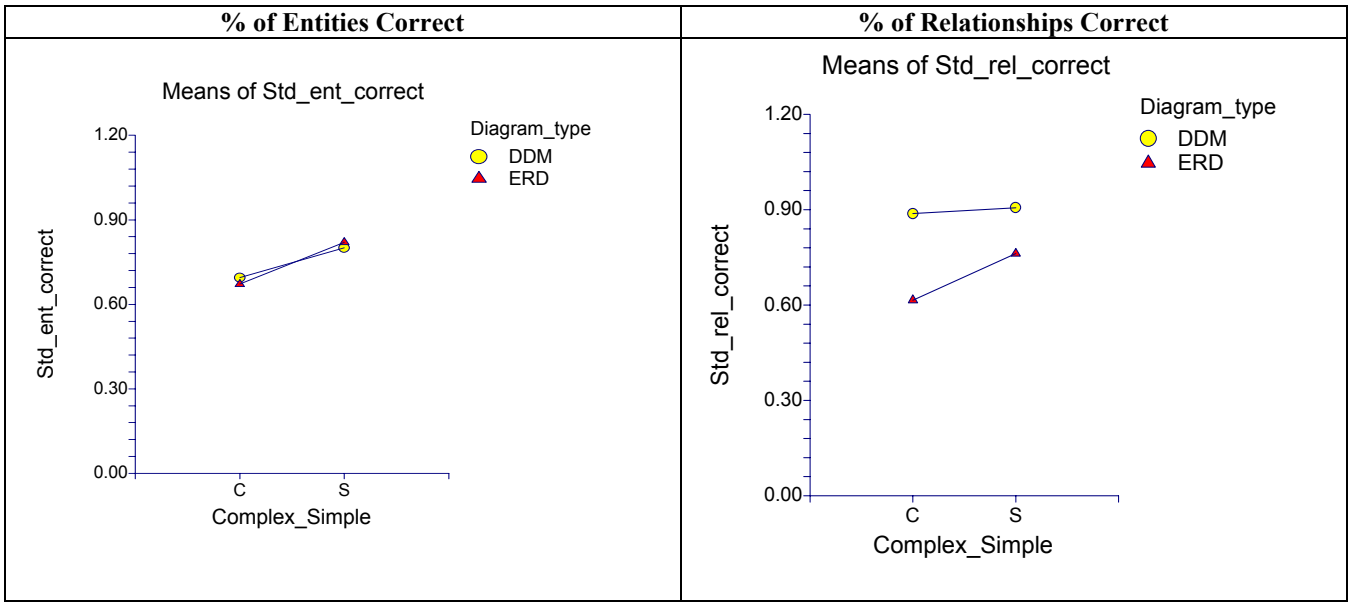

Figure 1. Graph of Results

References

Anderson, J. R. *Cognitive Psychology and Its Implications, 3rd ed.* New York: W.H. Freeman and Co., 1990

Business Intelligence/Data Warehousing Research Program. *Database Solutions III*. www.survey.com report, 2000.

Kimball, R. "A Dimensional Modeling Manifesto," *Database Magazine* (10:9), August 1997, pp. 59-68.

Mandler, G. "Organization and Memory," *Psychology of Learning and Motivation* (1), 1967, pp. 327-372.

Miller, G. A. "The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information," *Psychological Review* (63:2), March 1956, pp. 81-97.

Rolland, F. D. *The Essence of Databases.* London: Prentice Hall, 1998.

Weber, R. "Are Attributes Entities? A Study of Database Designers' Memory Structures," *Information Systems Research* (7:2), June 1996, pp. 137-162.