Chapter 7.5
Ex Ante Evaluations of Alternate Data Structures for End User Queries:
Theory and Experimental Test

Paul L. Bowen
The University of Queensland, Australia

Fiona H. Rohde
The University of Queensland, Australia

Jay Basford
The University of Queensland, Australia

ABSTRACT

The data structure of an information system can significantly impact the ability of end users to efficiently and effectively retrieve the information they need. This research develops a methodology for evaluating, ex ante, the relative desirability of alternative data structures for end user queries. This research theorizes that the data structure that yields the lowest weighted average complexity for a representative sample of information requests is the most desirable data structure for end user queries. The theory was tested in an experiment that compared queries from two different relational database schemas. As theorized, end users querying the data structure associated with the less complex queries performed better. Complexity was measured using three different Halstead metrics. Each of the three metrics provided excellent predictions of end user performance. This research supplies strong evidence that organizations can use complexity metrics to evaluate, ex ante, the desirability of alternate data structures. Organizations can use these evaluations to enhance the efficient and effective retrieval of information by creating data structures that minimize end user query complexity.

INTRODUCTION

In today's highly competitive business environments organizations are encouraging managers and other end users to query information repositories themselves (Delligatta & Umbaugh, 1993). The information requirements must be transformed into queries correctly; otherwise end users extract wrong data from the databases. The extraction of inappropriate data leads to poor decisions and unintended or dysfunctional organizational responses (Klein, 2002). The data structure of an information system can significantly impact the ability of end users to efficiently and effectively retrieve the information they need (Borthick et al., 2001). Whether users are interacting with databases, data marts, or data warehouses, the information retrieval processes can be improved by providing data structures tailored for end user query formulation tasks (Raisinghani, 2000).

In organizational settings, retrieved information is used to make operational, tactical, and strategic decisions. Numerous authors have written about the impact on decisions of information quality, particularly the accuracy of information (see, e.g., Hilton et al., 1981; Lorence, 2003; Strong et al., 1997; Wang & Strong, 1996).[1] Decision quality depends on the accuracy of the data retrieved (Fisher et al., 2003; Klein, 2002). Data retrieval accuracy depends, in turn, on two factors: first, the accuracy of the data stored in the base tables and, second, the accuracy of the data retrieved by users via their query formulations. The accuracy of data stored in base tables is enhanced by data structures that minimize data redundancy and maximize data integrity (Date, 2004; Hoffer et al., 2002). These internal schemas seek to maximize the accuracy of stored data, but are unlikely to be optimal for query purposes.

To minimize query errors, organizations need to identify, create, and maintain alternate external schemas for user queries, for example, create alternate user views. A single internal schema can provide the basis for an extremely large number of alternate external schemas. The external schema is the user level schema. Each user or group of users requires a view of the data structure tailored to their specific needs (Mannino, 2001). The end users' view (external schema) of the data is usually different from the manner in which the data are stored. From the end users' perspective their external view is "the data structure" and has the data needed for their requirements (Menzies & Hu, 2003). Different end users within the same end user group, for example, accounting or production planning, may have different requirements in terms of the subsets of attributes they need to perform their assigned tasks. The organization needs to determine which of the multiple potential external schemas is the most appropriate.

Minimizing the number of alternate external schemas for a particular group of users within an organization is important for at least three reasons. First, individuals require time and experience to achieve proficiency with a particular schema; that is, learning effects imply that individuals are likely to formulate queries more efficiently and effectively using fewer rather than a larger number of external schemas. Second, the greater the number of different external schemas users query, the greater the difficulty they will experience communicating with other stakeholders.[2] Third, maintaining each schema incurs processing, documentation, and training costs; that is, the greater the number of external schemas, the greater the maintenance costs.[3]

Because of the large number of potential external schemas and the need to minimize the number of external schemas supported by an organization, this study develops and empirically tests a method for evaluating, ex ante, the relative desirability of alternative data structures for user queries. The method is based on the theory that the data structure that yields the lowest aggregate complexity measure will be the data structure that produces the best end user query performance. In brief, the method consists of identifying alternative schemas that meet the end users' requests;