ABSTRACT

A research in progress concerning the problem of extending the logical independence approach to statistical query answering is described. The main ideas and limitations of a system which assists a user in obtaining a statistical table from a relational database by describing only elements that compose the table are discussed. The proposed system assists the user in formulating statistical queries by means of a universal relation interface, i.e., independently from the underlying database structure. To generate the result, an extended relational algebra is needed to process aggregations in queries function; heuristics for the interpretation of queries and their translation in terms of database primitives are the basis of the proposed approach.

1. INTRODUCTION

Statistical/scientific databases (SSDBs) are different from traditional databases in several aspects. Usually, they are more static (they represent consolidated events), i.e., update operations are infrequent and are applied to a small percentage of the data; on the other hand, a statistical query is, in general, more complex and involves a large amount of data [Shoshani 85]. Tipically, it is harder for the user to be formulate and more expensive for the system to evaluate.

There are two kinds of SSDB: micro and macro SSDBs [Wong 84]. Micro-SSDBs handle
information about single events or individual entities and their statistical features are embedded in the query system; on the other hand, macro-SSDBs contain mainly summary data of micro-SSDBs, i.e., the results of statistical operations such as cross tabulation or regression.

In this paper micro-SSDBs and queries to obtain statistical tables from a relational database are considered. Even if not all statistical queries are categorical in nature and not all statistics has to do with tabular output, we have restricted our attention to such cases of statistical analysis, since we consider this as a first step towards a more general approach to the problem.

The extensions of relational query languages proposed in the literature to handle statistical applications provide a formal environment for users that are sophisticated in terms of knowledge about the application domain and the task to be performed and expert in terms of the way in which the system works. In particular, it is required that users know: the data model, that is used to formalize the realm of interest, the syntax and semantics of the query language, and the way in which the application is represented in the system (the database scheme). In absence of one or more of the above requirements, query formulation becomes very hard.

For commercial database systems various interaction techniques have been proposed to take care of the way in which naive users work. One technique is based on the construction of database views. Views are designed and implemented by intermediary expert persons (application programmers), hence, they have a "static" nature, and are intended for regular users, whose interactions are predefined and frequently activated. In this way naive users can obtain personalized views on the database that are easier to understand and to use, but cannot modify easily their interaction goals or style. This does not seem to be suitable for statistical users of micro-databases that do not have, in general, a predefined task to perform and cannot interact with the model of the realm through an intermediary.

Another kind of interaction technique intended for naive users is the navigation technique implemented by browsers (see, for example, [Motro 86] and [Motro 88]) through the database. These tools are very simple to use, since they are based on a simplified data model and a set of elementary operations that move step-by-step the observation point in the database. Unfortunately, browsing is limited in its capability of performing complex tasks, and it is not clear how it can be used to specify a statistical query.

An intermediate interaction technique is query answering by finding "conceptual connections" between a set of user defined "goals" under the hypothesis that the database scheme is not shown to the user [Ausiello 86]. This paper concerns the problem of extending to the statistical environment a special case of this approach: the logical independence approach to query answering in relational databases, and in particular the possibility of using universal relation interfaces for statistical queries.

A universal relation interface [Ullman 83] allows the user to see a database as a single (universal) relation, by assuming that attributes have a global meaning, not limited to the relation