IDEA GROUP PUBLISHING # ITJ2248

701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

Providing Approximate Answers Using a Knowledge Abstraction Database

SOON-YOUNG HUH, Korea Advanced Institute of Science and Technology, Korea
JUNG-WHAN LEE, SK Telecom, Korea

As database users adopt a query language to obtain information from a database, a more intelligent query answering system is increasingly needed. Relational databases are exact in nature, but effectiveness of decision support would improve significantly if the query answering system returns approximate answers rather than a null information response when there is no matching data available. This paper proposes an abstraction hierarchy as a framework to practically derive such approximate answers from ordinary everyday databases. It provides a knowledge abstraction database to facilitate the approximate query answering. The knowledge abstraction database specifically adopts an abstraction approach to extract semantic data relationships from the underlying database, and uses a multi-level hierarchy for coupling multiple levels of abstraction knowledge and data values. In cooperation with the underlying database, the knowledge abstraction database allows the relaxation of query conditions so that the original query scope can be broadened and thus information approximate to exact answers can be obtained. Conceptually abstract queries can also be posed to provide a less rigid query interface. A prototype system has been implemented at KAIST and is being tested with a personnel database system to demonstrate the usefulness and practicality of the knowledge abstraction database in ordinary database systems.

INTRODUCTION

Query language is used as a handy tool to obtain information from a database. However, current query processing has not been satisfactory in supporting effective decision analysis. This is largely due to exactness in nature of the relational databases and the query languages. When there is no matching data available, database users usually get a null information response. In terms of rigidness of query structure, even expert users have been frequently frustrated by the precise query specification syntax which requires them to be well-versed with the database schema for formulating correct queries. For many queries, it may be better for the database system to produce approximate answers when no exact answer is available. Usability of the database queries enhances significantly if the users are allowed to write imprecise queries and the system understands the intent behind the queries. Cooperative query answering aims at developing such intelligent systems that can accept less-precisely specified queries, analyze the intentions of such queries, suggest related questions or relax query conditions, and provide approximate neighborhood answers (Chu & Chen, 1994, 1991; Cuppens & Demolombe, 1989; Minock & Chu, 1996; Motro, 1990; Prade & Testemale, 1984; Vrbsky & Liu, 1993). A variety of approaches have been attempted along with semantic distance, fuzzy set, rule, and conceptual classification.

The semantic distance approach (Ichikawa, 1986; Jain & Dubes, 1988; Motro, 1988) represents the degree of similarity between a pair of data objects by a numerical distance. This approach has advantages of ease and efficiency in developing query relaxation algorithms, since quantitative distances among data objects are easy to compute. Multiple distance metrics are also available for each attribute domain, and users are allowed to choose the direction of relaxation by mixing a selected set of distance metrics to compute a simplistic similarity value. However, this approach is limited due to the difficulty in transforming quantitative and qualitative data into a uniform quantitative measure. In addition, there is a lack of objective criteria for assessing semantic similarities among various data objects.

The fuzzy database approach (Buckles & Petry, 1982; Bordogna & Pasi, 1995; Bosc & Pivert 1995; Ram, 1995; Turban, 1988; Zemankova & Kandel, 1985) assumes that data objects in each domain can be assigned similarity values

between 0 and 1. In addition, the knowledge base can store various kinds of imprecise information such as mixed hair color (i.e., dark-brown: 0.6 brown and 0.4 black) and a certain person's residence (i.e., Boston or New York). Relations are extended to accept fuzzy values or to allow values that are sets of domain objects, and fuzzy comparators such as much-greater-than are provided to allow users to compose approximate queries. The answer to a query on a fuzzy database is also a fuzzy set, and thus a value in the answer has a membership degree between 0 and 1. On the basis of the rich semantic capabilities of the fuzzy set theory, the fuzzy database approach remedies the limitations of the relational database approaches by supporting various kinds of fuzziness derived from data themselves and linguistic queries. However, substantial semantics such as linguistic qualifiers for vague query (e.g., dark-brown), and impreciseness or similarity information on data should be previously incorporated in order to provide sufficient facilities for expressing various flexible data requests. Also, fuzzy database systems demand the development of their own query languages that fit in their fuzzy database model and application domain to accommodate fuzzy-based query conditions (e.g., much-greater-than), as well as crisp conditions.

The rule-based approach (Cholvy & Demolombe, 1986; Cuppens & Demolombe, 1989; Hemerly et al., 1994) adopts first-order logic as its formal framework and delineates semantic information about data objects and data integrity constraints using first-order formulas over a set of (base, built-in, derived) predicates. In this approach, the entire database is understood as a set of base predicates, and a database query also consists of a predicate rule whereby searching information is specified with free variables. The query is answered through conflict resolution and inference mechanisms, and query relaxation is carried out by coordinating the integrity constraints. The weakness of this approach includes a lack of systematic organization for guiding the query relaxation process and the less intuitive query answering process.

The conceptual clustering approach (Cai et al., 1993; Chen et al., 1990; Chu & Chen, 1994; Chu et al., 1991; Shum & Muntz, 1988; Vrbsky & Liu, 1993), to which this paper belongs, emphasizes the abstract representation of object instances, and replaces the instance attribute values by the abstract concepts to which they belong. For example in the PET attribute, dogs and cats are mammals, snakes and turtles are reptiles, and so on. The abstract representations form a clustering hierarchy where similar attribute values are clustered by the abstract concepts and the individual abstract concepts are related to one another by a certain abstraction function or the abstraction similarities. It aims at accepting a user-defined casual query that may be incompatible to the database schema, and transforming the query interactively into a schema-compatible form using the clustering hierarchy.

The approach helps the user to compose the query of interest with user's familiar expressions so that both user-friendly interface and vague queries are supported. Vague queries are accommodated and transformed by mapping appropriate query conditions or by relaxing the query conditions to be more generalized with a broader search scope. The conceptual clustering approach is specifically advantageous when the characteristics of data objects are qualitative and categorical (Chu & Chen, 1994; Chu et al., 1991). However, existing conceptual clustering approaches have limitations in the diversity of vague queries and the flexibility of clustering hierarchy maintenance. This is largely due to ambiguous distinction between value abstraction and domain abstraction information. Most studies focus on the construction of the clustering hierarchy itself from data object instances in the underlying database, but need further elaboration on the semantics of the hierarchy. If the value and domain abstraction information could be managed with explicit distinctions and support mechanisms, the clustering hierarchy can be constructed in a more systematic way by aligning a set of abstract values into a certain domain while maintaining value and domain abstraction information individually.

This paper proposes a practical conceptual clustering approach that can be easily developed for getting approximate answers from ordinary corporate databases. As a framework, it adopts an extended conceptual clustering framework, called knowledge abstraction hierarchy (KAH), capturing not only the value abstraction hierarchy but also the domain abstraction hierarchy and provides non-technical explanations for the approximate query answering mechanisms. Specifically, the value abstraction hierarchy consists of abstract values derived from specific data values in the underlying database and constitutes a hierarchical structure based on generalization and specialization relationships. The domain abstraction hierarchy is built on the data value domains and classifies them into super-domains and sub-domains. Based on the KAH, a knowledge abstraction database is designed to support approximate query answering mechanisms. A prototype system has been implemented at KAIST to demonstrate the usefulness of KAH in ordinary database application systems.

This paper is organized as follows. The next section discusses the knowledge abstraction hierarchy (KAH) in terms of two abstraction perspectives: value abstraction and domain abstraction. We construct a knowledge abstraction database that incorporates the KAH on a relational data model. The following section describes the approximate query answering operations on the basis of the knowledge abstraction database. The final section provides the conclusion of the paper.

CONSTRUCTS OF KNOWLEDGE ABSTRACTION HIERARCHY

Knowledge abstraction hierarchy (KAH) is a knowledge representation framework that facilitates multilevel