Relaxing Queries with Hierarchical Quantified Data Abstraction

Myung Keun Shin, Korea Advanced Institute of Science and Technology, South Korea
Soon Young Huh, Korea Advanced Institute of Science and Technology, South Korea
Donghyun Park, Inha University, South Korea
Wookey Lee, Inha University, South Korea

ABSTRACT

Query relaxation is one of the crucial components for approximate query answering. Query relaxation has extensively been investigated in terms of categorical data; few studies, however, have been effectively established for both numerical and categorical data. In this article, we develop a query relaxation method by exploiting hierarchical quantified data abstraction, and a novel method is proposed to quantify the semantic distances between the categorical data so that the query conditions for categorical data are effectively relaxed. We additionally introduce query relaxation algorithms to modify the approximate queries into ordinary queries, which are followed by a series of examples to represent the modification process. Our method outperformed the conventional approaches for the various combinations of complex queries with respect to the cost model and the number of child nodes.

Keywords: data abstraction; nearest-neighbor search; query relaxation; semantic distance

INTRODUCTION

In most database management systems, conventional query processing has focused on supplying exact answers to queries. There are a number of circumstances in which a user desires an approximate answer rather than an exact answer. For example, when a user does not always understand all about the data schema or when the queries contain errors syntactically or semantically, then the query results may be null or be given out too much. Then the user must amend or modify the query. An example in the data mining environment is when an initial query is answered, and that the answer can be considered as an anchor point from which the query can be relaxed to find more detailed information. Manual relaxation, however, for unsatisfactory queries is usually a drudgery and a time-consuming process that strongly requires a knowledge-based schema for the
database or data warehouse as well as query relaxation mechanism.

The query relaxation process can be explained in more detail by the following example. Consider an illustrative recruiting scenario in which the query that follows exists:

\[ Q: \text{Skill} = \text{‘COBOL’} \land \text{Salary} = \$40,000 \land \text{Age} = 40. \]

Assume that no result record comes out with the conventional query answering systems. Then in our approach, the first step to relax the query condition is as follows:

\[ Q_R: \text{Skill in (‘COBOL’, ‘C++’, ‘Java’)} \land 35,000 \leq \text{Salary} \leq 45,000 \land 37 \leq \text{Age} \leq 43. \]

Then, we sort the relaxed query results in terms of a distance measure between the original query and the objects, which will prove very useful for the applicants as they obtain a richer result of information. Finally, we get the results sorted by semantic distance \( D \), such as (a) < Abel, COBOL, $40,000, 40, \ D: 0.00 >, (b) < Albert, COBOL, $43,000, 40, \ D: 0.10 >, (c) < Harry, COBOL, $37,000, 38, \ D: 0.21 >, and (d) < Neal, C++, $38,000, 41, \ D: 0.39 >. In order to achieve this, a method of obtaining the approximate value and measuring the distance between the target value and the approximate value needs to be provided. For a numerical domain, such as salary and age, the difference between two values can be used as a semantic distance measure. For a categorical domain such as skill, the approximate values can be calculated by using a predetermined item distance table (Motro, 1990) or by the abstract hierarchy (Chen, Zhou, & Zhang, 2006; Chu, Yang, Chiang, Minock, Chow, & Larson, 1996).

The approaches based on the semantic distance approach (Motro, 1990; Muslea, 2004) use the notion of semantic distance to represent the degree of similarity between data values. Since query answering systems employing the semantic distance approach provide quantitative measures between target values and neighborhood values as a query result, users can retrieve approximate values more effectively using the measures as references to compare with different approximate values. However, for categorical data, the semantic distance approach has two problems because it employs a two-dimensional table to store distances among all pairs of data values. First, to find neighbor values of a target value, the system has to scan all the records related to the target value. Second, when a new value is added to a domain, it is required to consider distances between the value and all existing attribute values. This task contains a large amount of overhead to be done by a human operator, and moreover, human operators are liable to lose consistency in assigning distance data to a large number of values. In contrast, the approaches based on the abstraction hierarchy are suitable to dealing with categorical data. However, abstraction approaches cannot properly handle other data types, such as number, money, date and time, and so forth, and do not provide quantitative similarity measures among data values.

To overcome these problems, we propose hierarchical quantified data abstraction (HiQdA), which integrates the abstraction approach and semantic distance approach. The HiQdA uses the hierarchy structure of the abstraction approach and provides a quantitative measure between data values in the hierarchy. The abstraction hierarchy facilitates finding neighbor values for a target value quite easily. The distance information embedded in the HiQdA provides a more efficient method than the one based on a table. Maintenance of distance information due to the addition of a new value can be minimized since the change is localized in the hierarchy. This article will demonstrate how to calculate the similarity distance between two data values and introduce the query relaxation algorithm with HiQdA.

The rest of the article is organized as follows. The second section reviews prior related approaches. Then the article proposes the HiQdA as a new knowledge representation framework, and explains details of the query relaxation algorithm and examples using the
Applying Ontologies to Business and Systems Modelling Techniques and Perspectives: Lessons Learned
www.igi-global.com/article/applying-ontologies-business-systems-modelling/3312?camid=4v1a

Human Factors Studies of Database Query Languages: SQL as a Metric
www.igi-global.com/article/human-factors-studies-database-query/51074?camid=4v1a