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:

\[ Q_r: \text{Skill} \in \{\text{"COBOL"}, \text{"C++"}, \text{"Java"}\} \land 35,000 \leq \text{Salary} \leq 45,000 \land 37 \leq \text{Age} \leq 43. \]

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 as follows:

\begin{itemize}
  \item [(a)] < Abel, COBOL, $40,000, 40, \text{D: 0.00} >
  \item [(b)] < Albert, COBOL, $43,000, 40, \text{D: 0.10} >
  \item [(c)] < Harry, COBOL, $37,000, 38, \text{D: 0.21} >
  \item [(d)] < Neal, C++, $38,000, 41, \text{D: 0.39} >
\end{itemize}

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 neighbor-
An Overview of Ontology-Driven Data Integration