Supporting Imprecision in Database Systems

Ullas Nambiar
IBM India Research Lab, India

INTRODUCTION

A query against incomplete or imprecise data in a database, or a query whose search conditions are imprecise can both result in answers that do not satisfy the query completely. Such queries can be broadly termed as **imprecise queries**. Today’s database systems are designed largely for precise queries against a database of precise and complete data. Range queries (e.g., *Age BETWEEN 20 AND 30*) and disjunctive queries (e.g., *Name = “G. W. Bush” OR Name = “George Bush”*) do allow for some imprecision in queries. However, these extensions to precise queries are unable to completely capture the expressiveness of an imprecise query. Supporting imprecise queries (e.g., *Model like “Camry” and Price around “$15000”*) over databases necessitates a system that integrates a similarity search paradigm over structured and semi-structured data. Today’s relational database systems, as they are designed to support precise queries against precise data, use such precise access support mechanisms as indexing, hashing, and sorting. Such mechanisms are used for fast selective searches of records within a table and for joining two tables based on precise matching of values in join fields in the tables. The imprecise nature of the search conditions in queries will make such access mechanisms largely useless. Thus, supporting imprecise queries over existing databases would require adding support for imprecision within the query engine and meta-data management schemes like indexes.

Extending a database to support imprecise queries would involve changing the query processing and data storage models being used by the database. But, the fact that databases are generally used by other applications and therefore must retain their behaviour could become a key inhibitor to any technique that relies on modifying the database to enable support for imprecision. For example, changing an airline reservation database will necessitate changes to other connected systems including travel agency databases, partner airline databases etc. Even if the database is modifiable, we would still require a domain expert and/or end user to provide the necessary distance metrics and domain ontology. Domain ontologies do not exist for all possible domains and the ones that are available are far from being complete. Therefore, a feasible solution for answering imprecise queries should neither assume the ability to modify the properties of the database nor require users (both lay and expert) to provide much domain specific information.

BACKGROUND

The problem of supporting imprecise queries has already attracted considerable interest from researchers including those in fuzzy information systems (Morrisey, 1990), cooperative query answering and query generalization (Motro, 1998). More recent efforts have focused on supporting imprecise queries over relational databases by introducing ADTs (abstract data types) — for allowing storage and retrieval of non-standard data such as images, documents etc., and extending the query processor with functions for measuring similarity between tuples (Aditya et al, 2002; Ortega-Binderberger, 2003). Recently, work has been done on providing ranked answers to queries over a relational database (Bruno, Gravano and Marian, 2002). However, all the proposed approaches for answering imprecise queries require large amounts of domain specific information either pre-estimated or given by the user of the query. Unfortunately, such information is hard to elicit from the users. Further, some approaches require changing the data models and operators of the underlying database. A recent survey outlining challenges in integrating DB (database) and IR (information retrieval) technologies discusses the pros and cons of four possible alternatives for combining the two models (Chaudhuri, Ramakrishnan and Weikum, 2005).
Related Content

Heuristics in Medical Data Mining
[www.igi-global.com/chapter/heuristics-medical-data-mining/7780?camid=4v1a](www.igi-global.com/chapter/heuristics-medical-data-mining/7780?camid=4v1a)

Visual Data Mining for Discovering Association Rules
[www.igi-global.com/chapter/visual-data-mining-discovering-association/7751?camid=4v1a](www.igi-global.com/chapter/visual-data-mining-discovering-association/7751?camid=4v1a)

Association Rule Mining of Regional Data
[www.igi-global.com/chapter/association-rule-mining-regional-data/10568?camid=4v1a](www.igi-global.com/chapter/association-rule-mining-regional-data/10568?camid=4v1a)

Discovery of Protein Interaction Sites
[www.igi-global.com/chapter/discovery-protein-interaction-sites/10894?camid=4v1a](www.igi-global.com/chapter/discovery-protein-interaction-sites/10894?camid=4v1a)