Meaningful Keyword Search in Relational Databases with Large and Complex Schema

M. Kargar1, A. An1, N. Cercone1, P. Godfrey1, J. Szlichta2, X. Yu1

1York University, Toronto, Canada
2University of Ontario Institute of Technology, Oshawa, Canada

Introduction

1. Much of the world’s high-quality data remains under lock and key in relational databases.
2. Only users well versed in both SQL and in the schemas have access to the data.
3. Keyword search over relational databases was proposed a decade ago to offer an alternative way to query a database [1].
4. We present a new model for keyword queries in relational databases.
5. We propose schema-based ranking in order to rank final answers more effectively.
6. The experiments are conducted over the TPC-E database with a large and complex schema.

Motivation

An answer for the query “Arden Cynthia”. Such an interesting answer would be missed if we restrict the answer to be structurally minimal while covering all the query keywords. This happens in the previous work such as DISCOVER [1]. We do not miss this answer.

Framework Overview

1. The system identifies the database entities that are interesting to the user based on the keywords.
2. It returns the top minimal connected tuples that not only contain the query keywords but also cover the keyword roles (i.e., relations interesting to the user).
3. The system uses schema-based ranking that ranks the answers based on the minimal join trees that cover the keyword roles.

Ranking Example

Three possible trees (i.e., MJNSs, Minimal Joining Network of Schemas) for the query “Arden: company Cynthia: customer”.

Unlike Discover [1], we do not rank the answers based on their size but based on their relevance.

Discover ranks (a) as the best answer while our IF method ranks (b) as the best answer.

Evaluation

To have a ground truth, we construct a gold standard for relevance of answers from a transaction log of the TPC-E database. See [2] for description of our implemented demo.

References