Exploiting Latent Information in Databases via Database Embedding: technology, applications, ethics (Invited Talk)

Oded Shmueli
Technion – Israel Institute of Technology
Haifa, Israel
oshmu@cs.technion.ac.il

We are witnessing the emergence of AI-powered database systems, embedding AI-ideas and techniques in query processors, concurrency controllers, and more. We aim at improving relational querying, as well as other functionalities, by introducing another layer of data, word vectors, into traditional database systems. Word vectors originate in Natural Language Processing (NLP) where they are used to represent words in a language. In NLP, there are a number of methods for obtaining word vectors from text, we use a variation of one of these methods, word2vec.

The idea in a nutshell is as follows: we produce text from a relation (or a view thereof) and then use this text to generate a model, i.e., a set of vectors, for all terms in the database. Once the model is available, we can formulate Cognitive Intelligence (CI) queries. These queries may be realized by SQL queries, enhanced by User Defined Functions (UDFs) that take advantage of the model to formulate conditions that were previously practically not expressible in SQL.

The process of vector construction is different than in NLP. It reflects the characteristics of relations, with integrity constraints and named columns which contain various data types, strings, dates, numeric values, images and more. We call this process db2vec. There are a number of options for model generation: based on the textification of a single or multiple relations, incorporating external text sources (e.g., Wikipedia), incorporating externally produced models, standalone, or as building material for constructing a local model.

There are many application areas that may benefit from our approach: Commerce, Finance, HR, Science, and more. Whereas there are generic UDFs, some application areas require developing specialized UDFs. One example is a food database application in which a record has a list of ingredients, in decreasing order of importance.

A model reflects the textual and vector sources used to produce it. As decisions may be based on queries using the model, the production of models brings to the forefront issues of fairness and ethics. An important issue is the specifics of the data and text sources, their weighting in producing the model, and whether they are biased in some way.

Limiting information disclosure is also an important consideration. Especially within an organization, there is a need to share information. However, it would be desirable that this sharing enable productive work while hiding information that is not essential for that work. To this end, we introduce degrees of disclosure. Here, some information in the database is encrypted, some is simply not supplied, while additional information is intentionally supplied in the form of a model.