Representation Independent Database Analytics

Jose Picado
School of EECS, Oregon State University, Corvallis, OR
picadolj@oregonstate.edu

1. ABSTRACT

Over the last 20 years, users' information needs over structured data expanded from seeking exact answers to precise queries to finding entities or patterns *similar* to a given entity or pattern, discovering *interesting* entities and patterns, or predicting *novel* relations and concepts. As part of its response, the research community proposed a multitude of supervised and unsupervised algorithms to solve exploration and analytics problems over structured data in different contexts, such as similarity query processing, inexact pattern matching, and concepts and relationship prediction. Since the properties of interesting and desirable answers are no longer precisely defined in the query, and in many cases there is no query at all in the traditional sense, these algorithms use intuitively appealing heuristics to choose, from among all possible answers, those that are most interesting, important, and likely to satisfy the user's information need.

However, today's database exploration and analytics algorithms and tools are usable only by highly trained database analysts who can predict which algorithms are likely to be effective for particular representations of the underlying database. For example, given a relational database and training instances of a new target relation, (statistical) relational learning algorithms attempt to induce general (approximate) Datalog definitions of the target in terms of existing relations. Current (statistical) relational learning algorithms return different answers based on the schema chosen to organize the input database. For instance, they may learn accurate definitions over a highly normalized schema for a database and inaccurate ones over a less normalized schema for the same database or vice versa. As another example, researchers use deep learning methods to automatically learn useful representations for input knowledge-bases. It is shown that the results of some popular deep learning algorithms highly depend on the organization of the underlying database [1]. For instance, the representations learned over a database depend on the order of attributes in its relations.

Generally, there is no canonical representation for a particular set of content and people often represent the same information in different structures. Thus, users generally have to restructure their databases to some proper representations, in order to effectively use database analytics algorithms, i.e., deliver the insights that a domain expert would judge as relevant and important. To make matters worse, these algorithms do not normally offer any clear description of their desired representations and database analysts have to rely on their own expertise and/or do trial and error to find such representations. Nevertheless, we want our database analytics algorithms to be used by ordinary users, not just experts who know the internals of these algorithms. Further, the structure of large-scale databases constantly evolve, and we want to move away from the need for constant expert attention to keep exploration algorithms effective. More importantly, researchers often use analytics algorithms, such (statistical) relational learning and mining techniques, to solve other data management problems, such as data wrangling and transformation, entity resolution, and data integration. Thus, the representation dependence of data analytics algorithms make it difficult to provide usable systems for other data management tasks.

To cope with organizational heterogeneity and evolution in large-scale data, we propose a novel approach to database analytics that considers *representation independence*, i.e., the ability to deliver the same answers regardless of the choices of structure for organizing the data. We discuss our work on providing ordinary users with an arsenal of effective database analytics methods that are robust across multiple representations of the same information. We present our ongoing work on creating representation independent systems for (statistical) relational learning, deep learning, and data wrangling. In particular, we show how one can leverage traditional techniques in meta-data and schema management and query processing in database literature to design representation independent database analytics systems. Moreover, we discuss our findings on how to reduce the burden of users in hyper-parameter specifications over various representations.

2. REFERENCES

[1] R. Bailly, A. Bordes, and N. Usunier. Semantically Invariant Tensor Factorization. 2015.

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235