Representation Independent Data Analytics

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

1. ABSTRACT

Over the last years, users' information needs over structured data expanded from seeking exact answers to precise queries, to performing database analytics tasks, such as finding entities or patterns *similar* to a given entity or pattern, discovering *interesting patterns*, or predicting *novel relations and concepts*. As part of its response, the research community proposed a multitude of algorithms to solve exploration and analytics problems over structured data. Since the properties of interesting and desirable answers are no longer precisely defined in the query, these algorithms use intuitively appealing heuristics to choose, from among all possible answers, those that are most likely to satisfy the user's information need. Unfortunately, such heuristics typically depend on the precise choice of representation of the underlying database. Generally, there is no canonical representation for a particular set of content and people often represent the same information in different representations. Thus, in order to effectively use database analytics algorithms, users generally have to restructure their databases to some proper representation or change hyper-parameter settings. As a result, today's database exploration and analytics algorithms and tools are usable only by highly trained data scientists who can predict which algorithms are likely to be effective for particular representations of the underlying database, and under which settings.

To cope with the structural heterogeneity in large-scale data, we propose a novel approach to database analytics that considers representation as a first-class citizen. We introduce the concept of *representation independence* as the ability to deliver the same answers regardless of the choices of structure for organizing the data. Because representation independence may not always be achievable, we also consider *representation scalability*, the ability to return *similar* answers over different structures of 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 analytics systems over structured data, such as graphs and relational databases. 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/scalable database analytics systems. We now discuss some applications and challenges.

Relational learning: Given a relational database and training instances of a new target relation, relational learning algorithms attempt to induce general Datalog definitions of the target relation in terms of existing relations in the database. Current algorithms return different answers depending on the schema of the input database. For instance, they may learn accurate definitions over a normalized schema for a database and inaccurate ones over a less normalized schema for the same database or vice versa. Interestingly, this does not have to do with data content, but with data structure. Using the concepts of schema mappings, we develop a framework for defining the property of *schema independence* for relational learning algorithms. Current algorithms are generally not schema independent over different schemas because they explore different candidate definitions or in a different order. Further, current algorithms require that users know the schema to restrict the candidate definitions to explore. We propose an algorithm that employs database constraints to infer the best candidate definitions to explore and in which order, regardless of the schema.

Feature extraction: Another approach to perform analytics over structured data is to extract features from data, and then use these features as input to analytics algorithms. However, the values of some frequently used features change depending on the structure of data. A *robust feature* is one that does not suffer from variations in the structure of data. For instance, given a graph database, it is common to extract features such as indicators of centrality. These features are not robust because their value may change if the structure of the graph changes. We propose a bias/variance framework to measure the robustness of features, which helps users in extracting robust features from structured data.

Deep learning: Just as many fields are starting to use deep learning algorithms to learn features from raw data, it is interesting to consider using these algorithms over structured data. Recently, researchers have used deep learning algorithms to automatically learn useful representations for input knowledge-bases. However, it has been shown that the results of some popular deep learning algorithms highly depend on the organization of the underlying database. For instance, the representations learned over a database depend on the order of attributes in its relations.

8th Biennial Conference on Innovative Data Systems Research (CIDR '17) January 8-11, 2017 , Chaminade, California, USA.