Playing Information LEGO at Large Scale

Sebastian Michel
TU Kaiserslautern,Germany
smichel@cs.uni-kl.de

There is no such thing as a historical "big bang" where all data was created in one blink. Instead, information is constantly growing and evolving. New data is created and existing data enriched or adapted, over time—at high speed. This paper advocates research to handle dynamic information consisting of small pieces of data, say JSON documents, on-the-fly in order to constantly build-up new data on top of old. This requires efficient algorithms and pruning techniques to bring together matching information fragments in a scalable way, to do so incrementally over time, to avoid matching unrelated data, and to know when historic, although seemingly related data, is not anymore of interest for forthcoming matchmaking.

Motivation and Task Description: Consider an imaginary observer being placed in the middle of a "Big Data" stream of high pace, high volume, stemming from a multitude of independent sources. This observer notices users providing, for instance, during vacations critics, pictures, and other properties of restaurants. There is no single platform for collecting such information and users might simply give properties of visited restaurants by mentioning them in a Twitter tweet or Facebook post. While initially there might be only few information like address and name of the restaurant, with evolving time, more and more facts will show up. Take for instance Wikipedia, that serves as input to various knowledge bases, where facts about entities are virtually never complete, but get constantly enriched over time. In this process, new properties (attribute types as well as values) can be introduced at any time, by anyone. In the big data era, this observation is even more drastic as data originates virtually from everywhere; no central authority controls or at least oversees information creation and dissemination. As input, we can consider data being represented by generic JSON objects, i.e., a bag of possibly nested key-value pairs. Relating arbitrary *entity-centric information* that is "floating by" comprises two core tasks. First, it needs to be decided whether or not two or more pieces of information fit together, which means they describe the same entity and, thus, can be joined (aka. merged or matched). Second, how the incremental matchmaking over *several levels* of such merge processes can be handled efficiently over massive data streams. The former task is related to entity resolution, but would operate also on barely overlapping information fragments whose characteristics can greatly vary over time. There will be large amounts of newly arriving contents and historically merged information that need to be constantly revisited to check for validity and joinable information. Validity refers to data that was joined previously, but appears by current statistics invalid and should thus not be joined further-on.

Applications: There are various applications that could directly benefit from research conducted within the scope of this paper; primarily any area that can harness real-world, entity-centric information; for instance, data exploration and knowledge base creation and curation. With existing knowledge bases like Yago or DBpedia, there can be a cross-fertilization effect: on the one hand, we can harness knowledge bases to increase precision and recall of join decisions, and on the other hand, we can provide input to knowledge bases and further report on unlikely/problematic data combinations already present in them, for their curation.

Challenges: The vast scale and heterogeneity of data in absence of a narrow application scenario and intensive human labor to manually annotate/understand data, renders the ultimate goal described in this paper highly challenging. Specifically, we identified the following core challenges: (1) Semantic and structural heterogeneity of data: We need to determine/handle, for instance, ambivalent, previously not known, or cryptic attribute names and learn on-the-fly the hidden structure of entity-centric information that will eventually change over time again. To some extent entity names or attribute values in dictionaries and knowledge bases could help handling heterogeneity, but on the other hand, such information is often insufficient to capture the long tail of data. (2) Combinatoric explosion of potential matches: The more restrictive the matchmaking is done, the lower the fraction of determined true matches and the higher the precision. (3) Time varying results and data characteristics: Learning and adapting over time the obtained statistical models and understanding the validity interval (i.e., "expiration date") of previously matched data fragments is essential. (4) False positives/false negatives impacting forthcoming joins: Every wrong matchmaking is a threat to future matchmaking decisions that would incorporate faulty information or would miss matches due to having missed valid matches in the past.

Related Work: The scope of this paper tightly related to research on entity resolution, data integration, and (probabilistic) data stream joins. A first proof-of-concept implementation and more details on related work is presented in our recent workshop paper "Playing LEGO with JSON: Probabilistic joins over attribute-value fragments" published at KEYS@ICDE 2016.

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