NoSym: Non-Symbolic Databases for Data Decoupling

Souleiman Hasan
Insight Centre for Data Analytics, National University of Ireland, Galway
souleiman.hasan@insight-centre.org

1. INTRODUCTION

Under the Unique Name Assumption, users need to share agreements on signifiers to use in schema or data, e.g. to use "genre" and not "type" to refer to a movie's category. Agreements are difficult in open environments such as open data, and crowd-sourced databases, thus this assumption can be invalid. Schema matching and data integration can be limited in responding to this problem [1] as: (1) schemas might not be available a priori; (2) dataset-level schema/data mappings limit a user's ability to provide a contextual interpretation of a signifier suitable for a specific query-data matching task; and (3) data integration typically has an overhead which hinders the availability and low latency of databases.

2. THE PRINCIPLE OF DECOUPLING

Lack of shared agreements, or "decoupling", can be good. In Message-Oriented Middleware it has been leveraged to achieve scalability as parties do not have to: know each others' addresses (space decoupling), be active at the same time (time decoupling), or block each other (synchronization decoupling). We call the lack of shared agreements on the entities and schema dimension "data decoupling". From a semiotics perspective, the agreements needed on data are on a mapping between signifiers (schema labels or data values), and signifieds (what the labels and values mean). In databases, there is a little distinction between these two worlds which makes the problem intractable.

3. NON-SYMBOLIC DATABASES

We propose, as shown in Figure 1, a new paradigm which consists of three components:

(1) *Symbolic Encoder*: it translates data and queries from their symbolic form, e.g. English, into a non-symbolic representation. Deep learning models for example are trained on large corpora, e.g. Wikipedia, and can map a single word into a vector or tensor of hundreds of dimensions which can approximate its signified [3]. User can provide contextual data such as tags to tailor the encoding [2].

(2) *Non-Symbolic Database Management System*: it stores the tensors representations, index them, and process encoded queries against stored encoded data.

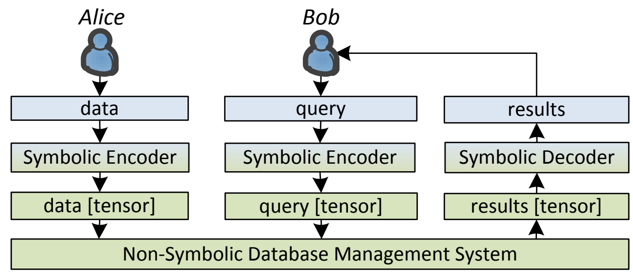

Figure 1: System components.

(3) *Symbolic Decoder*: it translates a query result into a symbolic form that the user can readily work with, e.g. English data rows.

4. CHALLENGES

Tensors are numeric and can be larger than the original symbols. We need formal and efficient models to store, index, and query them. Besides, non-symbolic tensors form a metric space where distance is defined with a topology that reflects the relatedness between the original values, e.g. "genre" and "type". This topology can be used for efficient storage and retrieval.

Encoders and decoders can become heterogeneous due to decoupling. However, as the NoSym database has native support for distances, it can natively catch similar or related signifiers. Besides, building agreements on encoders/decoders and their background corpora is not fine-grained, as with ontologies, and thus is more achievable. Furthermore, establishing a few anchor points between two metric spaces can lead to more points being automatically linked.

5. REFERENCES

[1] S. Hasan. *Loose Coupling in Heterogeneous Event-Based Systems via Approximate Semantic Matching and Dynamic Enrichment*. PhD thesis, National University of Ireland, Galway, 2016.

[2] S. Hasan and E. Curry. Thematic event processing. In *ACM/IFIP/USENIX Middleware*, 2014.

[3] T. Mikolov and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 2013.

This work has received funding under the European Commission's Seventh Framework Programme from ICT grant agreement no. 619660 (WATERNOMICS)