INRIA Saclay DirtyData project https://project.inria.fr/dirtydata

LIMSI,
bât 507, Orsay

Database models: embedding entities from their context

Research theme: machine learning, database, knowledge representation
Keywords: Distributional semantic, graph embeddings.
Duration & salary: 4 to 6 months, 500 € monthly
Research team: Parietal (INRIA Saclay)
Adviser: Gael Varoquaux, Alexandre Allauzen
Contact: gael.varoquaux@inria.fr
Application: Interested candidate should send CV and motivation letter

Context: Common *entities* are central to linking tables in databases. The notion of entity is a cornerstone of formal knowledge representation, in relational databases or in open-ended knowledge bases such as DBPedia. It is also important in information extraction from text. The discrete nature of entities sometimes come with limitations when manipulating data. The same entity may appear with different symbols: *eg* "Paris", "Parigi", and "Paris, Fr". The same symbol may denote different entities: "Paris" is also a city in Texas. “Orsay" is a different entity from "Paris", but it is close geographically. To represent those ambiguities and similarities, it is interesting to embed entities or symbols in a metric space, for instance representing each one in $\mathbb{R}^p$ [3].

In text processing, vectorial word embeddings represent links between words, such as related meaning, by similarities in $\mathbb{R}^p$ [4]. The core idea is that of *distributional semantic*: that the context of words informs on their meaning. It can be written in terms of a probabilistic language model [1]. Considering “sub-word” information, *ie* that observed strings for a symbol has internal structure, help linking different morphological variants of words [2].

In the semantic web, information is exposed via named entities explicitly linked together with an ontology, forming a graph of knowledge. Embedding the entities in a vector space where neighborhood reflects connectivity in the knowledge graph can also be useful to discover implicit links [5]. As in text processing, such embedding uses the context of entities to capture a meaningful representation, however the context is here defined by the graph.

Proposed work: The goal of this internship is to generalize these two approaches and develop symbol embedding for typical databases that are not as structured as a semantic-web knowledge base.

From a theoretical standpoint, it requires developing a model embedding symbols based on their (string) representation as well as a notion of context adapted to tables in a relational database.

From a practical standpoint, the models will be tested to embed symbols in from dumps of public databases downloaded from open-data portals of public-institutions: OECD, world bank, eurostats, data.gouv.fr, data.gov.uk. The embedding will enable finding commonalities between these datasets, to search or cross-analyze them.

Embedding elements of tables of relational databases can apply to either the symbols in the entries of the table or the column names, *ie*, the individual entities that the table represents, or the type of the entities in one column. Indeed, to come back to the example of *eg* "Paris", "Parigi", and "Paris, Fr", it can also be important to link a column name "ville" in one database to a column name "citta" in another. We will consider both cases, as they lead to a different notion of context, and hence different embedding strategies.

References

[1] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *JMLR*, 3:1137, 2003.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv:1607.04606*, 2016.

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, page 2787, 2013.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR Workshop Papers*. 2013.

[5] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29:2724, 2017.