Thesis Proposal:

The Role of Embeddings in Data-Driven Augmentation

Federico Dassereto

September 30, 2020

Abstract

Maecenas ipsum velit, consectetuer eu, lobortis ut, dictum at, dui. In rutrum. Sed ac dolor sit amet purus malesuada congue. In laoreet, magna id viverra tincidunt, sem odio bibendum justo, vel imperdiet sapien wisi sed libero. Suspendisse sagittis ultrices augue. Mauris metus. Nunc dapibus tortor vel mi dapibus sollicitudin. Etiam posuere lacus quis dolor. Praesent id justo in neque elementum ultrices. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. In convallis. Fusce suscipit libero eget elit. Praesent vitae arcu tempor neque lacinia pretium. Morbi imperdiet, mauris ac auctor dictum, nisl ligula egestas nulla, et sollicitudin sem purus in lacus. Praesent in mauris eu tortor porttitor accumsan. Mauris suscipit, ligula sit amet pharetra semper, nibh ante cursus purus, vel sagittis velit mauris vel metus. Aenean fermentum risus id tortor. Integer imperdiet lectus quis justo. Integer tempor. Vivamus ac urna vel leo pretium faucibus. Mauris elementum mauris vitae tortor. In dapibus augue non sapien. Aliquam ante. Curabitur bibendum justo non orci.

Morbi a metus. Phasellus enim erat, vestibulum vel, aliquam a, posuere eu, velit. Nullam sapien sem, ornare ac, nonummy non, lobortis a, enim. Nunc tincidunt ante vitae massa. Duis ante orci, molestie vitae, vehicula venenatis, tincidunt ac, pede. Nulla accumsan, elit sit amet varius semper, nulla mauris mollis quam, tempor suscipit diam nulla vel leo. Etiam commodo dui eget wisi. Donec iaculis gravida nulla. Donec quis nibh at felis congue commodo. Etiam bibendum elit eget erat.

Maecenas ipsum velit, consectetuer eu, lobortis ut, dictum at, dui. In rutrum. Sed ac dolor sit amet purus malesuada congue. In laoreet, magna id viverra tincidunt, sem odio bibendum justo, vel imperdiet sapien wisi sed libero. Suspendisse sagittis ultrices augue. Mauris metus. Nunc dapibus tortor vel mi dapibus sollicitudin. Etiam posuere lacus quis dolor. Praesent id justo in neque elementum

1 Motivation and Description

Since the rise of the World Wide Web, we have experienced an exponential growth in publicly available data. Data has become a crucial element in everyday life and many devices continuously produce and store them in a forest of different formats. This exponential growth has posed many challenges to scientists (even going so far as to create the figure of the *data scientist*) in every step in which data are traditionally used: from data collection to integration, from processing to visualization. The impact of *Big Data* has been

summarized in [9], where the key properties are **V**olume, **V**elocity, **V**ariety, **V**eracity and **V**alue.

In data science, it is increasingly the case that the main challenge is not in integrating known data, rather it is in finding the right data to solve a given data science problem. Today, data is a mass (uncountable) like dust, and data surrounds us like dust, even lovely structured data. Data is so cheap and easy to obtain that it is no longer important to always get the integration right and integrations are not static things (**V**elocity component). Data integration research has embraced and prospered by using approximation and machine learning. The uncontrolled nature of data manifests in large repositories of data (data lakes), in which both structured and unstructured data are stored. The peculiarity of data lakes lies in the fact that there is uncertainty about the presence of metadata describing the data themself. Furthermore, it is common the situation in which there is a lack of schemas, making traditional database approaches to integrating or querying data difficult to pursuit or even infeasible. Along with the uncertainty regarding the quality of the data, the amount of such data makes it infeasible also the traditional human-in-the-loop framework, since hand-labeling or manual rating of very large amounts of data is extremely expensive. It is easy to see that searching in data lakes is a complicated task, both in terms of time and methodology.

Following the path traced by Tim Berners-Lee et al. in [1] regarding Open Data publication and maintenance, many organizations are publishing data, augmentating tremendously the **V**olume and the **V**alue of such data. On the other hand, the explosion of sources providing data increases their **V**ariety, requiring new techniques to integrate data. At a processing level, the main difficult arisen by the multiple sources and possibly unstructured data is given by the interpretation, i.e., the semantic layer. To extract semantics from unstructured data, in the last years the concept of embedding has been proposed. Embeddings are predominantly used in learning representations of texts [10] and graphs [11]. In the last year, an approach exploting embedding to solve data integration task [2] has been published. This work can be seen as forerunner for the integration of embeddings in data management tasks, even if it works on tables from the same context.

A data scientist who is trying to solve machine learning tasks, frequently found herself in the situation in which there are not enough data to build a good model. Given the amount of data previously discussed, it would be very useful a framework which enables her to find automatically new features or samples to improve the model. Broadly, two main classes of augmentation can be distinguished: *Horizontal Augmentation*, in which new significant features are added to the dataset and *Vertical Augmentation*, in which new samples (tuples) are added to the dataset. Ideally, the two classes can be seen as the search for joinable (Horizontal) or unionable (Vertical) tables in the data lakes situation, and Knowledge Bases (KB) completion or extensions (vertical).

In this thesis, we focus on the problem of *using embeddings for automatic augmentation of data to improve machine learning tasks*. We believe that it is a crucial integration task that has not yet been explored, either as augmentation problem itself nor with the usage of embeddings. Few approaches that try to augment tables with respect to a repository have been proposed; the two closest approaches proposed until now show different problems: (i) a set of rules to decide if it is safe to avoid a join or not [8], restricted to a *pure relational* settings, i.e., the schemas of each table is known; and (ii) a feature selection algorithm working on a matrix in which the number of attributes is much larger than the number of tuples [4].

The goal of this thesis is to study how to stabilize the automatic increase and make it scalable to possibly huge tables or data lakes. The idea is to define indexing structures,

based on information theoretic measures, to make the search for joining tables fast and adapt to the variety of existing joining possibilities (one-to-one, one-to-many, many-to-one, many-to-many). More precisely, the thesis will proposed and index for horizontal augmentation on the data lakes scenario, as well as the integration of embeddings in the augmentation of KBs. Eventually, an embedding algorithm for KBs will summarize unified framework the two previous results.

Previous Works. From the above discussion, we would like to underline the three main topics of the proposal: *open data*, *data-driven augmentation* and *embeddings*. In order to familiarize with these three main concepts, we deepened them in the first year and published some works. Thanks to a long term collaboration with Prof. Michela Bertolotto (Univeristy College Dublin) and Laura Di Rocco (Northeastern University, Boston) started during my Master thesis entitled *Embeddings for Geospatial Ontologies Representation*, we were able to embed geographical ontologies onto a suitable space, on which we first evaluated its quality [5] and then its impact as query engine into a data-driven microblogs geolocation algorithm [7]. We further analyze the tuning of parameters to obtain higher quality embedding and summarized our results in [6]; the key idea is that a fine-grained tuning of the parameters could drastically improve the quality of the embedding in capturing semantic similarities.

On the other hand, we also worked in the Open Data domain, trying to expand an existing approach on Linked Open Data source selection [3] by introducing the concept of embedding to better associate context information. The work is in its conclusive part as the experimentation is going through and will be ready soon for submission. We remand to Section 4 for the discussion on the preliminary results obtained regarding the augmentation problem.

2 Reference Area and Relevance of Goals

The current proposal lies in the reference area "Data Augmentation to improve Machine Learning Tasks", with the aim of providing to a data scientist an augmented set of data which enables her to improve her model performances. The approach we propose differ from other existing solutions for the following aspects:

- Existing approaches mainly focus on predicting the uselfulness of a join under the realtional hypothesis, i.e., the schema informations are known. To this end, we plan to develop an approach that is agnostic to the knowledge of the schemas;
- Existing approaches which try to augment data against a repository perform very expensive joins, and evaluate thier work with respect to repositories of limited size; other existing approaches on repositories run features selection algorithms on a very large matrix derived by the join of all the joinable tables. To this end, we plan to develop an approaches that does not materialize any kind of join and is scalable to hundreds of thousands tables in a repository;
- Existing works on single tables identify the most relevant features in the table, according to a specific target. We plan to overcome to the single table assumption by indexing all the tables in a repository and identifying the most relevant features, and the relative tables, that improve the machine learning model performances;
- Other approaches simply return the set of joinable tables, usually with a treshold to allow for relaxed (imperfect) join, without any ranking. We plan to rank the returned tables according to the improvement that each of them will guarantee to the model.

3 State of the Art

In this section, we present related work that we consider relevant for the proposed research project. In order to facilitate reading, the discussion has been organized into fuor main sections, corresponding to the main concepts introduced in Section 1 and 2. Each of the four sections is in turn divided into sub-parts to further facilitate reading. The first one discusses embeddings approaches and their relevance to the proposal, the second introduces Open Data and their common operations, the third provides an overview of existing augmentation approaches and finally the fourth reviews functional dependencies discovery approaches.

3.1 Embeddings

3.2 Open Data

3.3 Augmentation Approaches

3.4 Functional Dependencies

4 Goals and Preliminaries

GOALS

- AGGIUNGI Embedding per ricerca in funzione di ML
- Studiare l'augmentation in termini di cosa esiste (entropia, FDs)
- Possibilità di augmenting verticale
- Introduzione delle KBs

5 Research Plan

Acknowledgements

So long, and thanks for all the fish.

References

[1] Christian Bizer, Tom Heath, and Tim Berners-Lee. "Linked data: The story so far". In: *Semantic services, interoperability and web applications: emerging concepts*. IGI Global, 2009, pp. 205–227.

[2] Riccardo Cappuzzo, Paolo Papotti, and Saravanan Thirumuruganathan. "Creating embeddings of heterogeneous relational datasets for data integration tasks". In: *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 2020, pp. 1335–1349.

[3] Barbara Catania, Giovanna Guerrini, and Beyza Yaman. "Exploiting Context and Quality for Linked Data Source Selection". In: *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. 2019, 2251–2258.

[4] Nadiia Chepurko et al. "ARDA: Automatic Relational Data Augmentation for Machine Learning". In: *arXiv preprint arXiv:2003.09758* (2020).

[5] Federico Dassereto et al. “Evaluating the effectiveness of embeddings in representing the structure of geospatial ontologies”. In: *The Annual International Conference on Geographic Information Science*. 2019, pp. 41–57.

[6] Federico Dassereto et al. “How to Tune Embedding Parameters in Geographical Ontologies”. In: *Under Submission.*. 2020.

[7] Laura Di Rocco et al. “Sherloc: a knowledge-driven algorithm for geolocating microblog messages at sub-city level”. In: *International Journal of Geographical Information Science* (2020), pp. 1–32.

[8] Arun Kumar et al. “To join or not to join? thinking twice about joins before feature selection”. In: *Proceedings of the 2016 International Conference on Management of Data*. 2016, pp. 19–34.

[9] Andrew McAfee et al. “Big data: the management revolution”. In: *Harvard business review* 90.10 (2012), pp. 60–68.

[10] Tomas Mikolov et al. “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781* (2013).

[11] Maximillian Nickel and Douwe Kiela. “Poincaré embeddings for learning hierarchical representations”. In: *Advances in neural information processing systems*. 2017, pp. 6338–6347.