ParHyTE: For Time-Aware Knowledge Graph Embeddings

Sruthi Gorantla (15190) [1]

Abstract

Stochastic Gradient Descent poses multiple challenges while training a model. It being inherently serial makes it difficult to parallelize without loss of computations. The main goal of this project is to exploit the sparsity while computing knowledge graph embeddings using translational embeddings. With theoretical guarantees, a parallelized version of SGD can still be used for TransE (Bordes et al., 2013) without losing the updates as shown in (Zhang et al., 2017). These guarantees can also be extended to a recent algorithm HyTE (Dasgupta et al., 2018), hyper plane based temporally aware knowledge graph embeddings. In this project, I show detailed analysis on how the sparsity of time information in *wiki_data* can be used to parallelize the computations of HyTE.

1. Introduction

1.1. SemMedDB

The SemMedDB (Kilicoglu et al., 2012) repository consists of information regarding semantic predications extracted from PubMed citations by preprocessing and stored for efficient access. The SemMedDB repository is implemented primarily as a MySQL relational database that consists of tables holding information regarding PubMed citations, relevant UMLS knowledge and the semantic predications. A brief description of the content of each table and the approximate number of records in it are provided in figure 1. We can construct a knowledge base in the form of triplets *(s, p, o)* from the *PREDICATION* table which has the fields *SUBJECT_CUI, PREDICATE, OBJECT_CUI*. This represents the knowledge graph of our study. A lot of additional information can also be mined from this database. Finally, a good representation of the entities in this graph will lay a path to calculate the semantic relatedness between them.

1.2. Knowledge Graph Embeddings

The history of knowledge graph embeddings goes back to the basic *TransE* (Bordes et al., 2013) algorithm where, with just the triplets information, knowledge graph embeddings are found by minimizing the following loss function.

$$L = d(s + p, o) \tag{1}$$

where $d(x, y)$ represents the Euclidean distance between x and y. This loss function makes sure that the subject, with respect to the predicate, is close in space to the object. These embeddings, however, donot take into account the independence of the types of relations between the entities. If the embeddings for the entities depend on the type of the relation between them, it makes more sense to compare the semantic relatedness between any two pairs of entities.

1.3. Distributed Computing

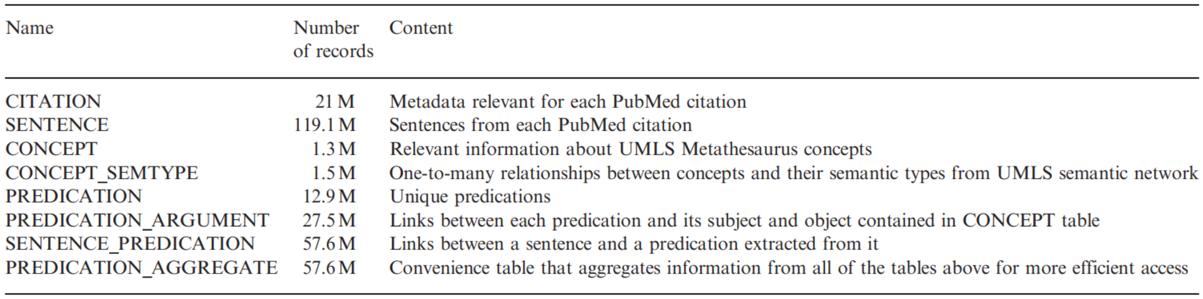
It is clear from the figure 1 that the number of such triplets that form semmeddb is very large. This necessitates the use of distributed computing in order to come up with the knowledge graph embeddings for the entire graph taking into consideration the extra information such as, co-occurrence counts, semantic types, relation types, text data etc. In this project, we aim at coming up with a novel architecture to compute the embeddings for the entities in semmeddb in a distributed manner. Further we aim at using the embeddings and the semantic relatedness between the entities in a downstream task to evaluate the embedddings.

2. Proposed Method

HEER (Shi et al., 2018) (**H**eterogeneous Information Network **E**mbedding via **E**dge **R**epresentations) is a recently proposed method to that effectively encodes the semantics of a knowledge graph with the use of edge representation. In this project, we plan to implement this model on semmeddb and further propose a new idea of using the extra information available in semmeddb. A future plan would be to use the learnt embeddings from semmeddb on a downstream task by proposing a method to find the semantic relatedness between the entities.

[1] Department of Computer Science and Automation, IISc Bangalore. Correspondence to: Sruthi Gorantla <gorantlas@iisc.ac.in>.

Course project for Machine Learning with Large Datasets, IISc, 2018.



Name	Number of records	Content
CITATION	21 M	Metadata relevant for each PubMed citation
SENTENCE	119.1 M	Sentences from each PubMed citation
CONCEPT	1.3 M	Relevant information about UMLS Metathesaurus concepts
CONCEPT_SEMTYPE	1.5 M	One-to-many relationships between concepts and their semantic types from UMLS semantic network
PREDICATION	12.9 M	Unique predications
PREDICATION_ARGUMENT	27.5 M	Links between each predication and its subject and object contained in CONCEPT table
SENTENCE_PREDICATION	57.6 M	Links between a sentence and a predication extracted from it
PREDICATION_AGGREGATE	57.6 M	Convenience table that aggregates information from all of the tables above for more efficient access

Figure 1. The list of tables in SemMedDB

References

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. Translating embeddings for modeling multi-relational data. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 26*, pp. 2787–2795. Curran Associates, Inc., 2013.

Dasgupta, S. S., Ray, S. N., and Talukdar, P. Hyte: Hyperplane-based temporally aware knowledge graph embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2001–2011. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/D18-1225.

Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., and Rindflesch, T. C. Semmeddb: a pubmed-scale repository of biomedical semantic predications. *Bioinformatics*, 28 (23):31583160, Aug 2012. doi: 10.1093/bioinformatics/bts591.

Shi, Y., Zhu, Q., Guo, F., Zhang, C., and Han, J. Easing embedding learning by comprehensive transcription of heterogeneous information networks. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining - KDD 18*, 2018. doi: 10.1145/3219819.3220006.

Zhang, D., Li, M., Jia, Y., and Wang, Y. Efficient parallel translating embedding for knowledge graph. In *IEEE/WIC/ACM International Conference on Web Intelligence 2017 (WI 2017)*, 2017.