7th International Conference on Computational Social Science IC^{2}S^2
July 27-31, 2021, ETH Zürich, Switzerland

Embedding Heterogeneous Hierarchical Structures

Sundong Kim

Keywords: Hierarchical Embedding, Heterogeneous Information, Representation Learning, Korean Districts and Businesses Embedding, **LocEmb**

This paper introduces a practical way to learn representations of heterogeneous concepts in the same hyperbolic space when each concept lies in latent hierarchical structures. The proposed tree-integrated method plays the role of tying heterogeneous trees together by referring to a concept map between trees, allow to get embeddings of different concepts in the same space. A possible implication of this technique includes a recommendation system for estate agencies to suggest which district their customers should consider relocating to maintain the lifestyle they enjoyed. Figure 1 shows the approach to finding representations from heterogeneous concepts.

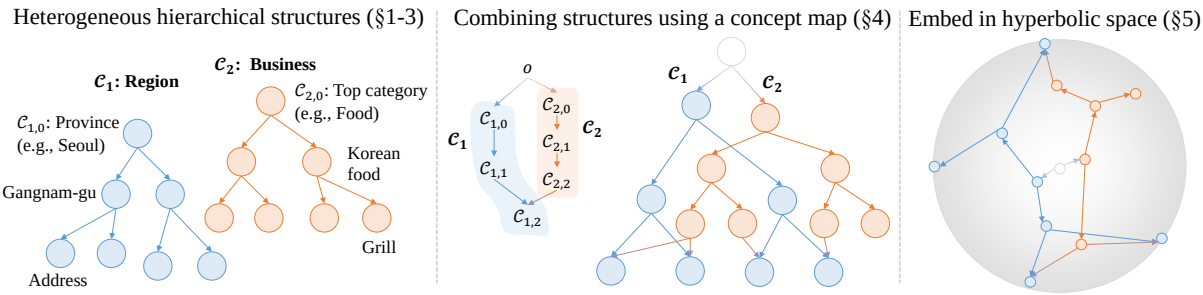

Figure 1: Illustration of how heterogeneous hierarchical structures are embedded together.

1. *Pick entities*: Among all entities in the data, decide which entities to embed. In real data, often, a variable contains multiple entities with hierarchies. (e.g., A variable *address* can be divided into various levels of identity, including city, district, road name, building number.)
2. *Categorize entities*: Categorize entities according to the concept they belong to. (e.g., city, district, road name, and building number belong to *region* concept. Restaurant, bar, cocktail bar, and the name of the bar belongs to *business* concept.)
3. *Prepare trees*: If there is a hierarchical relationship between entities, prepare a hierarchical database—tree, for each concept. Multiple trees can be generated from the source data.
4. *Combine trees*: By defining the dummy center node o as a parent of all the root nodes of each category $\mathscr{C}$, multiple individual tree structures can be merged into a unified tree with root o. Relationships between the nodes can be further specified by making additional connections if any node pair has a parent-child relationship. In this paper, I call this extra parent-child relationship as *concept map*. The structure of the concept map can be determined by the nature of the data and the modeler's domain knowledge. Concept maps can also be automatically made using knowledge graphs. Figure 1 illustrates an example of combining two trees using a concept map. A blue tree $\mathscr{C}_1$ representing regions, and an orange tree $\mathscr{C}_2$ representing businesses are integrated using a concept map.
5. *Embed together*: By projecting the combined tree in a Poincaré ball with Riemannian optimization [1], the model can learn the representation of each entity in hyperbolic space.

(a) Lodgings around Baekdun-ro 133 beon-gil in Gapyeong, the embedding of the street has the highest similarity with the embedding of a business category "accommodation."

(b) Bars around Yeonsei-ro 7 an-gil in Seoul, the embedding of the street has the highest similarity with the embedding of a business category "Hofbräuhaus."

Figure 2: Validation by querying street names on a search engine. In fact, there are many shops around the road that has the highest similarity to a particular business.

Case Study: Location and Business Embedding The algorithm is applied to the public commercial real estate dataset provided by the division of Small Enterprise And Market Service (SEMAS) in the Ministry of Small and Medium-sized Enterprises and Start-ups, South Korea. The dataset released in Dec 2019 is used and the most recent dataset can be retrieved from the website (`https://www.data.go.kr/dataset/15012005/fileData.do`). The data is tabular form, including the information of off-line commercial real-estates with their addresses and business type. As a result, 100-dim of embeddings are generated, comprising $\mathscr{C}_1$—9,000 *districts*, 110,000 *roads*, and $\mathscr{C}_2$—840 *business categories*, 1,482,860 *businesses* in Korea. Embedding results are released, with the name of LocEmb. Embedding results can be found in the LocEmb project repository (`https://github.com/seondong/locemb`).

Performance Analysis Intra-concept and cross-concept similarity analysis is performed to measure the performance of LocEmb. As an example, Figure 2 validates the superiority of our embeddings by confirming the results from a commercial search engine that the streets with the highest cosine similarity to specific business types (lodging, bar) were actually the most popular areas for those businesses. Extensive analysis results can be found on the web article on the repository (`https://github.com/Seondong/LocEmb/blob/master/LocEmb-EDA.ipynb`).

Conclusion This work presents an embedding approach when the data lies in multiple hierarchies. Through tree integration, different concepts can be embedded in the same hyperbolic space. The effectiveness of the approach is proved by similarity analysis within a concept and cross-analysis between two concepts. We released LocEmb, million-scale pre-trained embeddings for location and business in South Korea to facilitate social science research. LocEmb will be served as the initial resource for modeling point-of-interest recommender systems [2], forecasting real-estate prices, understanding spatial polarization, simulating the effect of social-distancing policies in Korea during COVID-19 epidemics.

[1] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," in *Neural Information Processing Systems*, 2017.

[2] B. Kim, S. Kim, S. Kim, and J.-G. Lee, "Customer revisit prediction using macroscale mobility information," in *Korea Software Congress*, 2019.