Embedding Heterogeneous Hierarchical Structures

Sundong Kim
Data Science Group, Institute for Basic Science
Daejeon 34126, Republic of Korea
sdkim0211@gmail.com

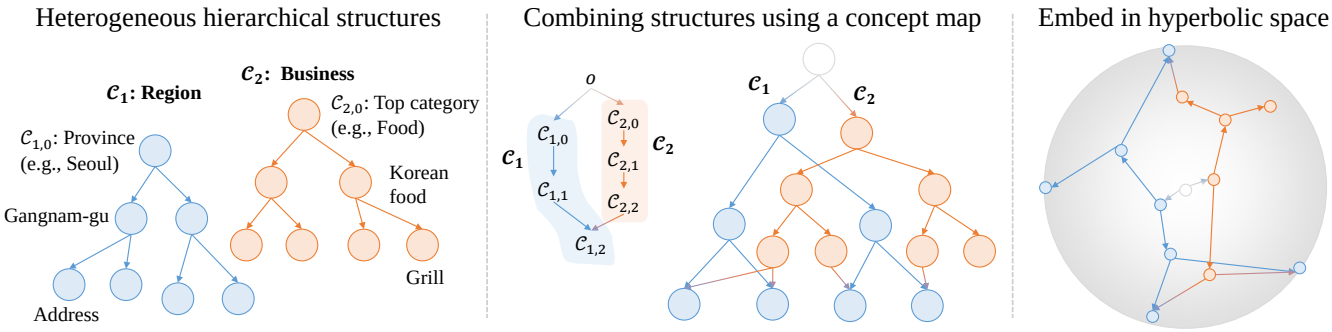


Figure 1: Illustration of how heterogeneous hierarchical structures are embedded together.

ABSTRACT

This paper introduces a practical way to learn representations of heterogeneous concepts in a 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 apply Poincaré embedding to find representations of diverse entities. The biggest advantage of this method is to get embeddings of different concepts in the same space. The embedding results LocEmb proved the effectiveness of the method by using commercial real estate datasets covering 9,000 districts and 2.7 million businesses in Korea.

KEYWORDS

Hierarchical Embedding, Heterogeneous Information, Representation Learning, Released Embeddings of Korean Businesses: LocEmb

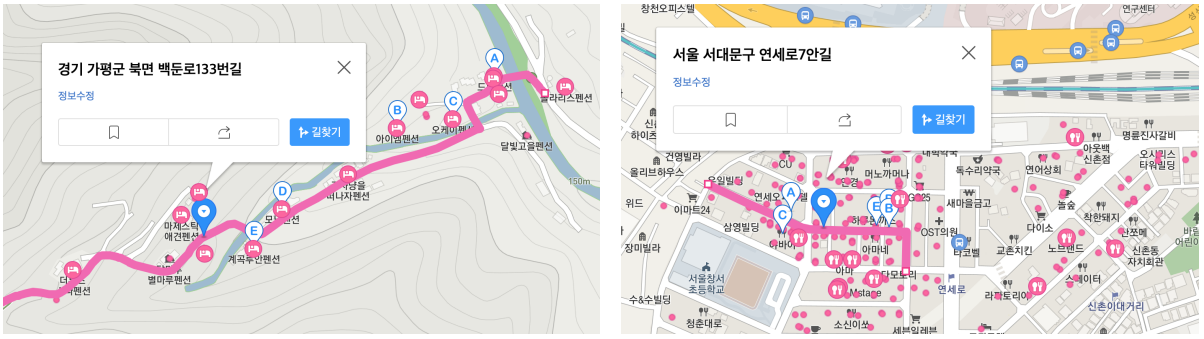
1 EMBEDDING MULTIPLE HIERARCHIES

How to develop a practical but effective technique that maps the representation of each entity to the most appropriate embedding space when multiple heterogeneous hierarchical structures exist? In this paper, I introduce a tree integration method that allow to take advantage of Poincaré embedding into heterogeneous trees. With this method, we can train an hyperbolic embedding model with the following steps:

- *Pick entities*: Among all entities in the data, decide which entities to embed. In the real data, often a feature contains multiple entities. (e.g., An address can be divided into various levels of identity, including city, district, road name, building number, etc.)
- *Categorize entities*: Categorize entities according to the concept they belong to. (e.g., City, district, road name, and building number belongs to spatial category. Restaurant, bar, cocktail bar, and the name of the bar belongs to business category.)
- *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.

At this stage, modelers can decide to obtain the embedding of each entity by mapping each tree in different hyperbolic spaces through the Poincaré method [1]. The obtained hyperbolic embedding is known as a superb initialization point for a follow-up model. However, the embedding obtained in this stage can be sub-optimal due to the lack of convenience, especially when niche interpretation is required. Since the hierarchical representations are trained separately, entity embeddings generated from each category cannot be comparable to each other. If the two entities belong to different categories, their representations also exist in different subspaces. As a result, it is difficult to analyze the similarity between entities that belong to different categories. To overcome these weaknesses, entity representation from different category must be located in the same hyperbolic space. This idea naturally leads me to develop a tree integration as a follow-up step:

- *Combine trees*: By defining the dummy center node o as a parent of all the root node of each category 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, this extra parent-child relationship is called as *concept map*. The structure of the concept map can be determined by the nature of the data and the domain knowledge of the modeler. Using the latest technique from knowledge graph mining, it can be also possible to generate concept maps automatically. Figure 1 illustrates an example of combining two trees using a concept map. A blue tree C_1 representing regions and an orange tree C_2 representing businesses are integrated using a concept map. After integration, we can obtain joint embeddings.



(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.

- *Embed together*: By projecting the combined tree in a Poincaré ball with Riemannian optimization, the representation of each entity can be learned in hyperbolic space.

2 LOCATION AND BUSINESS EMBEDDING

2.1 Used Dataset

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[1]. 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 840 *business categories*, 9,000 *districts*, 110,000 *roads*, and 1,482,860 *businesses* in Korea. Embedding results are released, with the name of LocEmb.

2.2 Summary of the LocEmb Embedding

Embedding results can be found in the LocEmb project repository.[2] First, location entities from the first tree C_1 are as follows:

- `Provincial`, `Municipal-level`: Total 251 high-level districts including province, (metropolitan) city, and some of their autonomous districts named as `Gu`.
- `Submunicipal-level`: Total 8,587 districts including two types of districts: 5,005 `Beopjeong-dong` and 3,582 `Haengjeong-dong`.
- `Road-level`: Total 110,722 roads including boulevards, avenues, and streets.
- `Address-level`: Total 1,979,166 addresses including 991,559 road name addresses and their older 987,607 land-lot addresses.

Next, business entities from the second tree C_2 are as follows:

- `Wide-level`: Total 9 categories.
- `Middle-level`: Total 94 categories.
- `Narrow-level`: Total 737 categories.
- `Individual-businesses-level`: Total 1,802,617 businesses including franchises and its branch name. There are 1,482,860 businesses with unique name and 319,757 branches of franchises. Each address in C_1 maps to each business in C_2.

Note: More details on administrative divisions of South Korea can be found on this article[3]. The concept map that I used is similar to the one shown in Figure 1, but a little more complicated.

2.3 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 showing 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 our project page.

3 CONCLUSION

In this work, I present an embedding approach when the data lies in multiple hierarchies. Through tree integration, different concepts can be embedded in the same hyperbolic space. To see the practical effect, we introduce LocEmb, embeddings for location and business embeddings in South Korea. We show the effectiveness of the proposed approach by performing similarity analysis within a concept and cross-analysis between different entities.

ACKNOWLEDGMENTS

I would like to thank Beomyoung Kim for the discussion during his internship. Although I finally use the public data, I thank Loplat[4] for supporting their dataset to get this project started.

REFERENCES

[1] Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *NeurIPS*. 6338–6347.

[1] https://www.data.go.kr/dataset/15012005/fileData.do
[2] https://github.com/seondong/locemb
[3] https://en.wikipedia.org/wiki/Administrative_divisions_of_South_Korea
[4] https://loplat.com/