Joinable Tabular Data Search with Table Embedding

Yoonseok Choi , Han-joon Kim *

Department of Electrical and Computer Engineering, University of Seoul, Korea

choiys8819@naver.com, khj@uos.ac.kr

Abstract. We can generate new meaningful information by integrating two or more relational tables into one table. However, finding tables that can be joined within a large dataset space (such as Data Lake, and Data Space) is a very labor-intensive and time-consuming task. In this paper, we propose a new technique (called *n-Join-pair*) of discoverying three or more table datasets that can be joined by using the previously proposed *CNE-Join* technique.
Keywords: Deep Learning, Data Integration, Join, Embedding, Relational Table.

1 Introduction

By performing a join operation on a specific column for two or more relational tables, it is possible to obtain meaningful information that was not known before joining. There have been many related studies, including CNE-Join [1], which automatically finds two tables that can be joined when a large number of table data is given. However, as far as we know, no research has yet been conducted to find three or more tables that can be joined when there is a lot of tabular data and to rank the joined tables according to their quality. This paper proposes a new joinable table dataset discovery technique called *n-Join-pair*, which finds three or more relational tables that can be joined with the quality information.

2 Proposed Method

Fig. 1 shows the execution process of *n-Join-pair*. The CNE-Join is a joinable dataset discovery technique that considers both column name embedding and named entity recognition to automatically find two tables that can be joined. Basically, we first utilize the CNE-Join method to discover a set of possible join pairs for two tables, with which we can discover three or more joinable table dataset. For example, if tables A and B can be joined on columns a and b, respectively, and tables A and C can be joined on columns a and c, respectively, then tables A, B, and C is possible to join on columns a, b, and c. Next, our method maps both the joined table and its source tables into the same embedding vector space through table embedding [2]. Here, it is necessary to evaluate the quality of the joined table, and for this, we calculate the average distance between the embedding vectors of the joined table and its source tables. Our *n-Join-pair*

* Corresponding Author: Han-joon Kim (khj@uos.ac.kr)

** This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2022R1A2C1011937).

technique takes multiple tables as input and ultimately produces a ranked set of joined tables. To obtain more meaningful ranking information, the average distance between the vectors of the joined table and its source tables is normalized and converted to a value between 0 and 1. Then, the range from 0 to 1 was divided into 10 sections, and the average value of the values included in the joined table for each section was calculated. For performance evaluation of the proposed method, we have conducted diverse experiments using 50 relational tables collected from Kaggle. Among the significant experimental results, Fig. 2 shows the average containment values according to the distance between the joined table and its source tables. As can be seen in this figure, as the average distance in the embedding space increases, the containment value of each joined table tends to decrease. The Containment formula for three source tables of joined table is defined as Eq. (1), and is used to quantitatively evaluate the quality of the joined table. The containment equation used in Eq (1) is a modification of the containment equation in the paper in [3]. The denominator in Eq. (1) refers to the minimum number of unique values in each of join columns when joining three tables, and the numerator refers to the number of unique values in the join column after join operation. A lower containment value means more information loss after the join operation, and so we try to obtain join results with a high containment value. As a result, this measure can be used to perform a ranking task on the joined results for *n* tables.

$$\text{Containment (a, b, c)} = \frac{|a \cap b \cap c|}{\text{MIN}(|a|, |b|, |c|)} \quad \textbf{(1)}$$

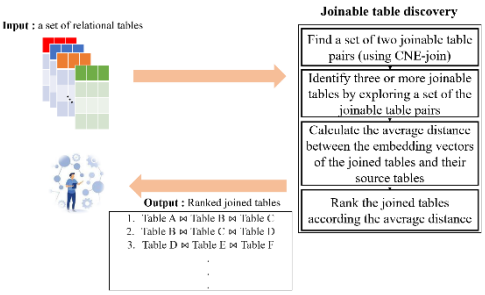

Fig. 2. Process of discovering n-join pairs

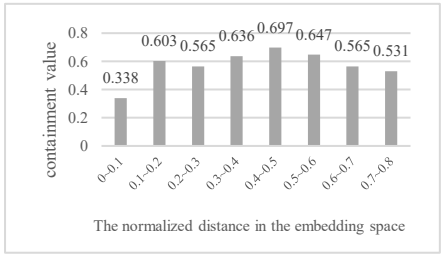

Fig. 1. The average containment values according to the distance between the joined table and its source tables in the same table embedding space.

References

1. Yoonseok Choi, Han-joon Kim.: Joinable Dataset Discovery with Integrating Column Embedding and Named Entity Recognition. In: Korea Artificial Intelligence Conference, pp. 118-120. (2023)
2. Jong-chan Yoon, Han-joon Kim.: A Relational Table Embedding Technique for Data Fusion. The Journal of Society for e-Business Studies, vol. 27, no. 3, pp. 1-19. (2022)
3. Erkang Zhu, Fatemeh Nargesian, Ken Q. Pu, and Renée J. Miller.: LSH ensemble: internet-scale domain search. Proc. VLDB Endow. 9, 12, pp. 1185-1196. (2016)