cases

doc_1		doc_2		decision	id
authors	• Yuyang Dong • Kunihiro Takeoka • Chuan Xiao • Masafumi Oyamada	**authors**	• Yuyang Dong • Kunihiro Takeoka • Chuan Xiao • M. Oyamada	DUPLICATES	295
title	Efficient Joinable Table Discovery in Data Lakes: A High-Dimensional Similarity-Based Approach	**title**	Efficient Joinable Table Discovery in Data Lakes: A High-Dimensional Similarity-Based Approach		
publication_date	2020-10-26 01:39:35+00:00	**publication_date**	2020-10-26 00:00:00		
source	SupportedSources.ARXIV	**source**	SupportedSources.SEMANTIC_SCHOLAR		
journal	None	**journal**			
volume		**volume**			
doi		**doi**	10.1109/ICDE51399.2021.00046		
urls	• http://arxiv.org/pdf/2010.13273v4 • http://arxiv.org/abs/2010.13273v4 • http://arxiv.org/pdf/2010.13273v4	**urls**	• https://www.semanticscholar.org/paper/5bca4dc4084c685a512ec129876d99095d83fa65		
id	id-8575887719954632018	**id**	id-8721838215039899188		
abstract	Finding joinable tables in data lakes is key procedure in many applications such as data integration, data augmentation, data analysis, and data market. Traditional approaches that find equi-joinable tables are unable to deal with misspellings and different formats, nor do they capture any semantic joins. In this paper, we propose PEXESO, a framework for joinable table discovery in data lakes. We embed textual values as high-dimensional vectors and join columns under similarity predicates on high-dimensional vectors, hence to address the limitations of equi-join approaches and identify more meaningful results. To efficiently find joinable tables with similarity, we propose a block-and-verify method that utilizes pivot-based filtering. A partitioning technique is developed to cope with the case when the data lake is large and the index cannot fit in main memory. An experimental evaluation on real datasets shows that our solution identifies substantially more tables than equi-joins and outperforms other similarity-based options, and the join results are useful in data enrichment for machine learning tasks. The experiments also demonstrate the efficiency of the proposed method.	**abstract**	Finding joinable tables in data lakes is key procedure in many applications such as data integration, data augmentation, data analysis, and data market. Traditional approaches that find equi-joinable tables are unable to deal with misspellings and different formats, nor do they capture any semantic joins. In this paper, we propose PEXESO, a framework for joinable table discovery in data lakes. We target the case when textual values are embedded as high-dimensional vectors and columns are joined upon similarity predicates on high-dimensional vectors, hence to address the limitations of equi-join approaches and identify more meaningful results. To efficiently find joinable tables with similarity, we propose a block-and-verify method that utilizes pivot-based filtering. A partitioning technique is developed to cope with the case when the data lake is large and cannot fit in main memory. An experimental evaluation on real datasets shows that our solution identifies substantially more tables than equi-joins and outperforms other similarity-based options, and the join results are useful in data enrichment for machine learning tasks. The experiments also demonstrate the efficiency of the proposed method.		
versions		**versions**			