Madhav Sigdel

Data Warehouse

05/01/2020

I found this arXiv publications dataset on website called figshare . Here is the link.

<https://figshare.com/articles/arXiv_publications_dataset_with_simulated_citation_relationships/6449756/1>

Dataset is posted on 06/06/2018 by[Jacek Miecznikowski](https://figshare.com/authors/Jacek_Miecznikowski/5344169)[,Dominik Tomaszuk](https://figshare.com/authors/Dominik_Tomaszuk/4376236). This collection contains data aquired from arXiv.org via OAI2 protocol.

Data was created through pseudo-random simulation to evaluate scientific research impact using six popular author-level metrics.

I would like study research paper on science. So, to get the information about the publication, author research subject and time of research, I have chosen this dataset. Dataset contains five tables

I choose subsets of Publication, Categories and Authors table for my project work. Publication and Categories have insert, update and delete anomaly problems. I remove these problems by 1st Normalization. Further, I did 2nd and 3rd Normalization by using primary and foreign key constraint and created relational database with five tables.