Microsoft - DAT278x

From Graph to Knowledge Graph – Algorithms and Applications

LAB 2 – Knowledge Graph

Heterogeneous Knowledge Graph Exploration – Instructions

1. Objective

In this lab, we will familiarize you with the contents we have covered in module 4 and 5 about basic knowledge graph properties by

- constructing a heterogeneous academic graph from a subset of the Microsoft Academic Graph (MAG); and
- exploring the created heterogeneous graph to examine its properties.

2. Input files

There are two input files for this Lab, they are the same two files used in the environment setup smoke test.

The first input file describes Paper-Author relations [**Paper_authors.tsv**]; and is the same with Lab 1 input. For its details, please refer to Lab 1 instruction file.

The second input file describes Paper-Venue relations [**Paper_venue.tsv**].

- It contains a set of academic publications; each publication is associated with one venue at which it is published.
- The file is formatted as tab-separated value fields with two columns:
    - the first column represents the paper Id; and
    - the second column represents its publishing venue name.

For example, below is a sample input of the paper-venue file:

PaperId	VenueName
2125526935	IJCAI
107169386	IJCAI
2742657268	KDD

Please note:

A. The publishing venue could be either journals or conferences, in this lab session, for illustration purpose, we only pick a small set of computer science conferences as venues to limit the data size.
B. In the released Microsoft Academic Graph, the Paper-Venue relationship is more complex as the Venue is represented as VenueId (foreign keyed to a separate Journal or Conference table), instead of the string format VenueName. We use a simplified schema version to help the students easier understand the core relationship. We also hope it make better sense to the audience if they are familiar with these top computer science conference names.
C. Each paper may have multiple authors, but it is associated with only one venue. [We also simplify and do NOT consider multiple-venue scenario in this Lab.]

3. Output graph – Heterogenous author-paper-venue Graph

What is an heterogenous author-paper-venue graph?

It is a heterogenous graph with three types of nodes and three types of edges (no weights associated with edges).

- **Node**:
  - Author
  - Paper
  - Venue
- **Edge**:
  - Paper-Author
  - Paper-Venue
  - Author-Venue

A ***triple*** is linked among one author, one paper, and one venue if and only if the author publishes this paper in this venue.

The constructed output heterogeneous graph can be represented as a list of ***triples***. For example, one sample output is listed below:

AuthorId	PaperId	VenueName
112778	2125526935	IJCAI
1152630	107169386	IJCAI
757571	2742657268	KDD

4. How to Run the Lab

4.1 There is an associated U-SQL script *DAT278x_Lab2_KnowledgeGraph.usql.* Please download and save it.

4.2 Please follow the same steps in the Challenge Lab Setup Instruction, on how to upload the input file (Paper_venue.tsv) into the ADL folder (course-kg/input/) – step 4.

The full input file path is "/course-kg/input/Paper_venue.tsv". **[Please note:** this is the **SAME** file that we used for Smoke Test run, you do **NOT** have to do it again if you passed the Smoke Test.**]**

Please verify that the "course-kg/input/" folder shall contain both the "Paper_venue.tsv" file and the "Paper_authors.tsv" file to be ready to run Lab 2 script.

4.3 Please follow the similar steps in step 5 from the Challenge Lab Setup Instruction, open the downloaded *DAT278x_Lab2_KnowledgeGraph.usql.*

4.4 Same as in Lab 1, you are expected to fill in the missing code in *DAT278x_Lab2_KnowledgeGraph.usql* script to answer the questions. The missing code is surrounded with some reminder comments like: "//Q1: START CODE HERE" and "//Q1: END CODE HERE" – if we use Question 1 as an example.

4.5 Same as in Lab 1, for the missing code section, the missing part could be:

- *4.5.1* either provided directly in the question statement, so that you can copy and paste the code to the script; or
- *4.5.2* there are multiple choices for the candidate missing code snippets, please pick the right one which can answer the stated question; copy and paste the selected code to the script's expected position.

4.6 After filling in the missing code lines for all questions, you can follow the Challenge Lab Setup Instruction – step 5.3 and 5.4, to run the script and check the results. The heterogeneous graph will be constructed and the answers to the questions will be in the outputting file: *Results_Lab2_KnowledgeGraph.tsv*. Please use the answers in this file to finish all the multiple choices questions as well as numerical input questions in **Challenge Lab 2 on Knowledge Graph – Questions** section.

Same as in Lab 1, you can choose to run the script and answer questions either all at once, or section by section with multiple script runs. Please be careful on how to comment out sections and pay attention to the logical and data dependency between code blocks.

5. More questions to think about

5.1 What other interesting insights/analysis you can generate from the constructed heterogenous graph? Can you give some examples and use U-SQL queries to generate results? For example, who is the most productive author (has the largest number of publications) for a given venue (e.g. “KDD”)?

5.2 Which venue are the most “similar” ones to a given venue (e.g. “KDD”)? Let’s define the “similarity” between two venues as the number of common authors who published in both venues. Can you write U-SQL scripts to generate the top 5 most similar venues to “KDD”?

We hope you enjoy this lab and have learnt more about the academic heterogenous graph construction and properties!