UNIVERSITE LIBRE DE BRUXELLES
Faculté des Sciences Appliquées
Academic Year 2020–2021

January 20, 2021

INFO-H-415 – Advanced Databases
Esteban Zimányi
Final Project

DBLP[1] is an online bibliographical database for computer science containing more than 5 million references. Its content is publically available in XML format[2]. The DBLP collection follows the BibTeX format and contains several types of references such as article, inproceedings, proceedings, book, incollection, phdthesis, masters-thesis, and www. Various fields describe the above types of references such as author, editor, title, booktitle, pages, year, address, journal, volume, number, month, url, ee, cite, publisher, note, crossref, isbn, series, school, and chapter. Notice that the not all fields are allowed in all reference types; please refer to the DTD file for this information.

You need to load this dataset into the graph database Neo4J. There are many possible ways to perform this task, one of them is given by the GraphDBLP project[3] and the related Github repository[4]. Due to the size of the dataset, which has more than 500 MB when compressed, it is recommended to use a small excerpt of this data to start developing this project. When everything is working you can envision to use the full dataset and compare the loading time and the performance of the queries (see below). Excerpts of the dataset can be found on the web[5] but you can extract your own.

After loading the dataset in Neo4J, you must implement in Cypher 20 analytical queries. The queries should be of different classes, aiming at describing various aspects of the dynamics of the domain of academic publications. Examples of queries are given next (you can include these among the 20 queries you will implement).

1. Give the number of publications for each type.
2. Give the name of authors.
3. Give the names of authors who are also editors.
4. Give the authors ordered by the number of publications, in descending order.
5. Give the authors and the number of publications for each type.
6. Give the author(s) having the highest number of publications.
7. Give for each author the total number of publications and the number of publications by type.
8. Give the list of proceedings that have at least one editor that is also author of at least one article in the proceedings.
9. Give for each author the number of co-authors and the number of joint publications with each of them.
10. Give the distance of author "Hector Garcia-Molina" with respect to other authors. Two authors that write together a publication have distance 0. If an author a write a publication with author b and if author b write a publication with author c, then a is at distance 1 from c if a and c have not published together.

Remarks

- The project can be done in groups of 2 students, the same that did the semester project.
- You can use your preferred programming language and tools for loading the dataset in Neo4J.
- The project must be submited through the UV web site on January 31, 2021 at the lastest. The submission will contain a report in PDF format explaining in particular your approach to load the dataset and the queries. It will also contain a zip file all the code needed to reproduce your solution from the input XML file.

---

[1] `http://www.informatik.uni-trier.de/~ley/db/`
[2] `http://dblp.uni-trier.de/xml/`
[3] `https://dblp.uni-trier.de/search?q=GraphDBLP`
[4] `https://github.com/fabiomercorio/GraphDBLP`
[5] For example, `https://hpi.de/naumann/projects/repeatability/datasets/dblp-dataset.html`