Analyzing Relative Incompleteness of Movie Descriptions in the Web of Data: A Case Study

Wancheng Yuan, Elena Demidova, Stefan Dietze, Xuan Zhou

CHALLENGE

In the context of Linked Open Data (LOD), datasets are published or updated frequently, constantly changing the landscape of the Linked Data Cloud. In this case study we investigate relative incompleteness among subgraphs of DBpedia, Freebase, and LinkedMDB and propose measures for relative data incompleteness in LOD. The study provides insights into the level of accuracy and actual conflicts between different LOD datasets in the movie domain.

A CASE STUDY

We investigate the impact of the neighborhood size to better understand the reliability of cross-dataset links. The relative incompleteness in the DBpedia/Freebase pair reaches 100% in the first order and 89% in the second order, meaning that all the Movie entities in the datasets are affected by incompleteness issues. The overall 1st Order Incompleteness of the Movie entities in the other dataset pairs is also pretty high, e.g., 70% for LMDB/Freebase and 56% for LMDB/DBpedia.

DISCUSSION

Incompleteness is a very common phenomenon in these datasets, and its number increases significantly with increase of investigated entity neighborhood. The main causes of relative incompleteness observed during our experiment are due to different interpretations of properties in the datasets.

OUTLOOK

Our method of classification and identification of quality issues provides not only insights into the level of agreement between datasets but also into the overall quality of datasets. In future work we intend to extend these approaches to infer knowledge about the correctness and agreement of schemas.