Structured Language Model: Keyword Query Disambiguation

Prachi Jain, Indian Institute of Technology Delhi
ADVISORS: Arash Termehchy, Prof. Marianne Winslett

Background

Document retrieval models for text-centric, semi-structured documents cannot be applied to data-centric semi-structured data sets, since their structure contains rich semantic and structural information that sheds light on the meaning of their content. A term may not necessarily refer to the same concept throughout the document.

Goals

The keyword query interface must determine the information needs of the user on the basis of the query and the database only, with no training data available.

Research Plan

- Compute the probability of an entity generating a query.
- Compute the probability of various mappings given the query and entity.
- Use the model to compute the odds of relevance of the entity given the query.
- Implement and add the script to the current system.
- Assess the results of the standard INEX queries.

Terminology

The term “Godfather” appears in different attributes in the above database fragment, but only one satisfies the user’s information need.

Research Results

Our structured language model is found to deliver better ranking quality than the current KQIs do.

Fundamental Questions/Challenges

- How to determine the importance of the entity’s attribute type in each candidate answer.
- How to determine the relevance of an entity set in a set of entity sets.
- How to compare potential entity results from different entity sets.