Ultra-heterogeneous Entity Graphs

Large and complex graphs capturing millions of entities and billions of relationships between entities.

- Freebase: 1.9 billion triples
- DBpedia: 3 billion triples
- YAGO: 120 million triples
- Linked Open Data: 52 billion triples

Applications: search, recommendation systems, business intelligence, health informatics, fact checking

Need for a Quick Overview

Approach 1: Schema Graph

Schema Graph itself can be too complex.

Approach 2: Schema Summary

- Schema summarization in relational database
  [Yang PVLDB09, Yang PVLDB11]
- XML summarization [Yu VLDB06]
- Graph summarization [Tian SIGMOD08, Zhang ICDE10]

Schema graph of “Film” domain in Freebase
- Entity graph: 2M entities, 18 M edges
- Schema graph: 63 entity types, 136 edges

Optimal Preview Discovery

Find the preview with highest score that satisfies:
- Size constraint
- Number of key attributes $K$
- Distance between two preview tables $d$

Concise preview, dynamic programming algorithm

We assume all $K$ key attributes are ordered arbitrarily.
- optimal concise preview (k, n, X) is the best of:
  - optimal concise preview (k, n, X-1)
  - optimal concise preview (k, n, X-1, X)
  - $X$ key attributes with $n$ non-key attributes

Tight/Diverse preview, Apriori property algorithm

1. Construct $2^k$ cliques by enumerating all attribute pairs
2. For $i = 3$ to $k$, generate $i$-cliques based on Apriori property
3. Find the $k$-clique with highest score, return as optimal preview

Algorithms

Aggregate Scoring

Score of the Preview

$$f_0 + \sum_{k, n, X} f_{k, n, X}$$

Attribute Scoring

- Key attribute scoring
- Coverage-based method: Coverage[Film] = $M$
- Random walk-based method: Stationary distribution of a random walk process defined over the schema graph
- Non-key attribute scoring
- Coverage-based method: Coverage[Genre] = $S$

User Study

Domains: film, books, music, TV, people
- Hand-crafted preview tables: 62 PhD students in Database research group
  - Individually and as a group
  - 50 $\times$ gift card

Existence/experience questions
- Schema graph
- Concise preview
- Tight preview
- Diverse preview
- Foundation ground truth

Comparison between rankings by our approach and the crowd,
Pearson Correlation Coefficient (PCC)