Background

XML schemas can be very large and complicated, making it hard for new users or programmers to understand the schema well enough to issue queries or work with the database.

Clustering

- We attempt to summarize the schema by clustering together leaves that are related to each other.
- This is done by clustering the complete graph on the leaves with an appropriate distance metric.

Distance Metrics

- Inverse of mutual information between the leaves calculated on the basis of the content of the leaves in the database; this captures different entities present in the dataset.
- TF-IDF vector space model distance. This captures different subjects.

Results for IMDB Dataset

Vector space model distance works best because of the textual nature of data.
E.g., if asked for 7 clusters for IMDB:

Results for DBLP Dataset

Mutual information does well because of presence of distinct entities. E.g., if asked for 5 clusters for DBLP: