

Schema Discovery in RDF Datasets

CONTEXT: Linked Data on the Web

- Structureless
- Incomplete type information
- Noisy

GOAL: Automatic Schema Extraction

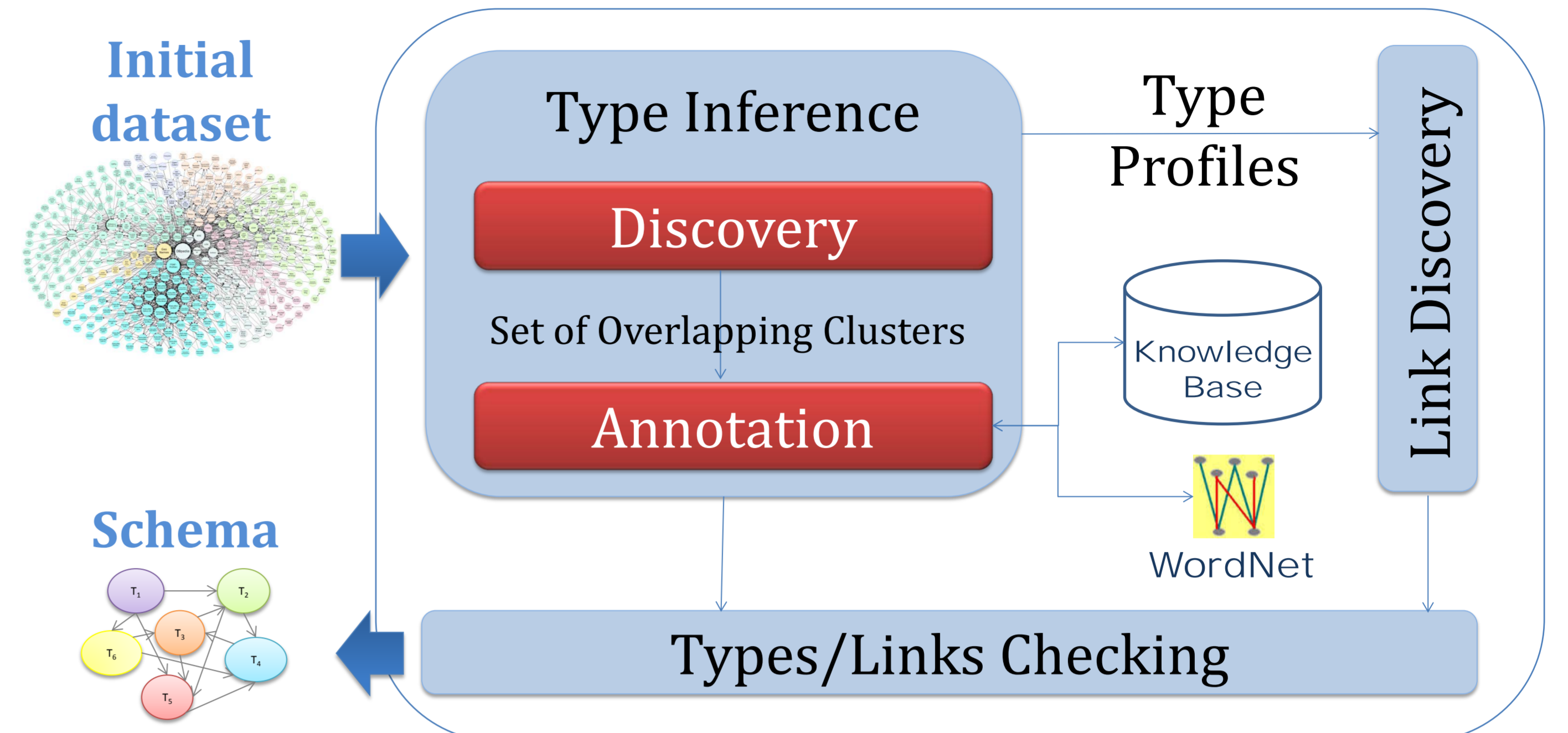
- Types
- Semantic links
- Hierarchical links

TYPE DISCOVERY

- Density-based clustering
 - Groups data according to the neighbors density
 - Robust to noise and deterministic
 - Detects classes of arbitrary shape
 - No required number of classes
- Automatic detection of similarity threshold
 - According to the density distribution of the dataset
- Type profiles
 - Type properties with their frequencies
 - Probability for an instance of a type to have a property
Example: « Conference »
- Overlapping types
 - Identification of the important properties for the types
 - Comparison of type profiles

$\langle \text{URL}, 0.5 \rangle, \langle \text{date}, 1 \rangle, \langle \text{made}, 1 \rangle, \langle \text{isHeldAt}, 1 \rangle, \langle \text{authorList}, 1 \rangle$

SCHEMA EXTRACTION PROCESS



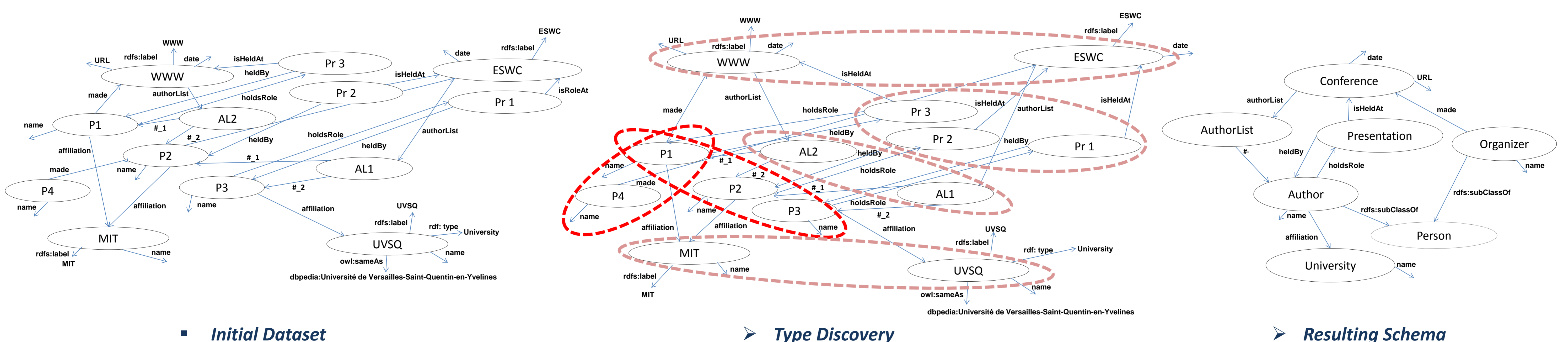
LINK DISCOVERY

- Semantic links
 - Inferred from user-defined properties in the initial dataset
 - Considering the direction of properties in the type profiles
- Hierarchical links
 - Hierarchical clustering of type profiles

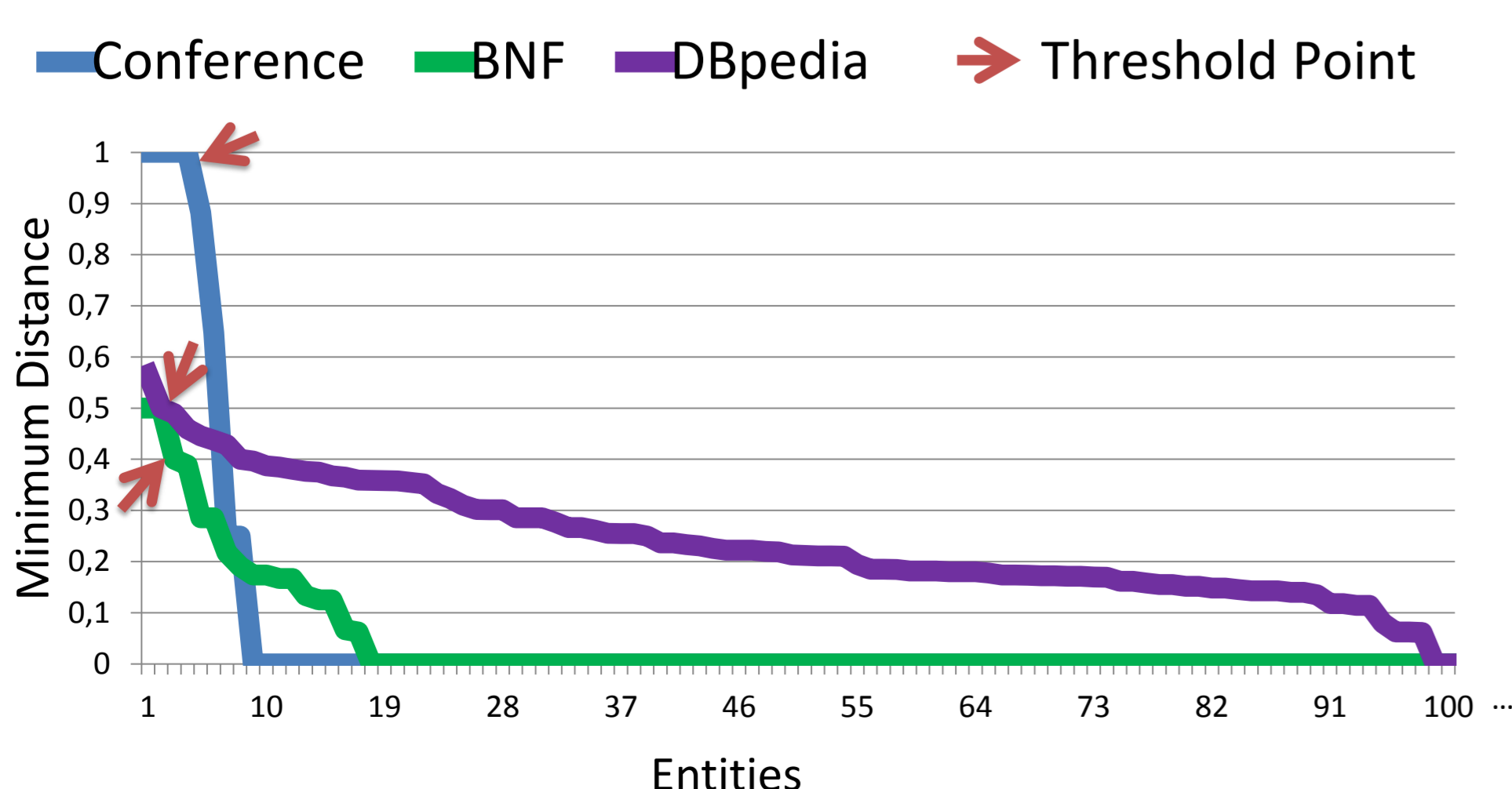
TYPE ANNOTATION

- Types extracted from a knowledge base
- Name based annotation
 - Searching for the types of resources having the same value of the name property
- Property based annotation
 - Searching for the types of resources having the same properties than the type profile, using WordNet for terminological conflicts
- Vocabulary based annotation
 - Searching for the domain/range of the properties in a standard vocabulary

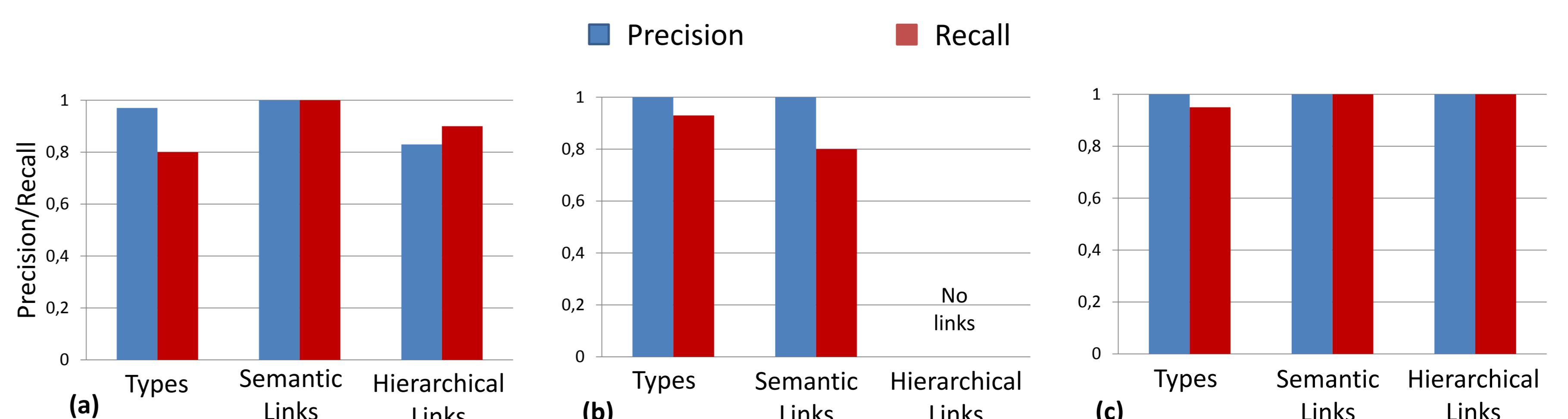
USE CASE



EXPERIMENTATION



Automatic detection of similarity threshold



Evaluation of schema discovery quality in Conference (a) BNF (b) and DBpedia (c) datasets