Interleaving Clustering of Classes and Properties for Disambiguating Linked Data

Takahiro Komamizu, Toshiyuki Amagasa, Hiroyuki Kitagawa (University of Tsukuba, Japan)

Linked Data

Link together and **Open** to public

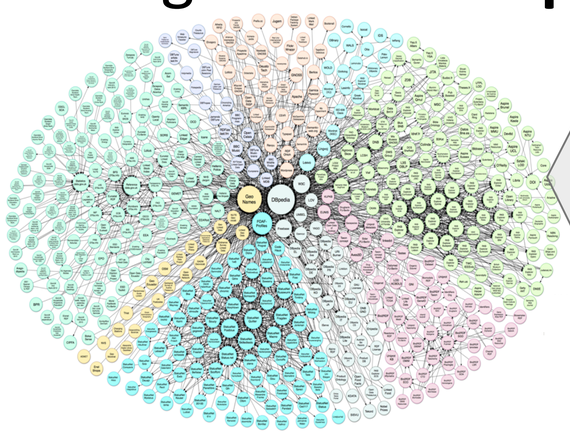
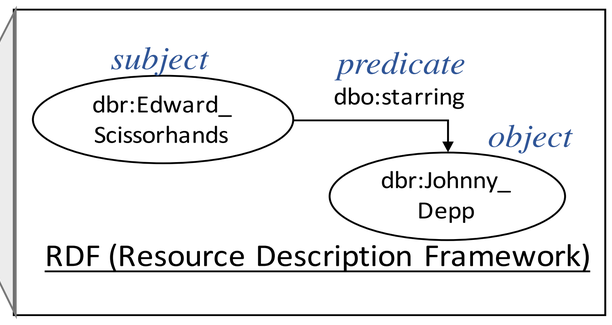


*Image from http://lod-cloud.net/




```
select ?movie
where{
    ?movie rdf:type dbo:Film;
    dbp:starring dbr:Johnny_Depp.
}
```

SPARQL

Ambiguity Problem

Class ambiguity

- Similar classes with different URIs
- e.g., foaf:Person and dbo:Person

Property ambiguity

- Similar properties with different URIs
- e.g., dbp:starring and dbo:starring

- Inappropriate SPARQL queries for users
- Undesired burden on adding new entities

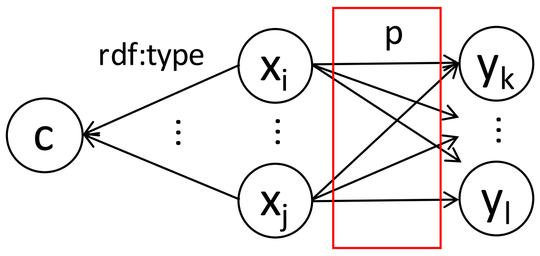
Proposed Approach: CPClustering

Basic idea: clustering onto classes and properties

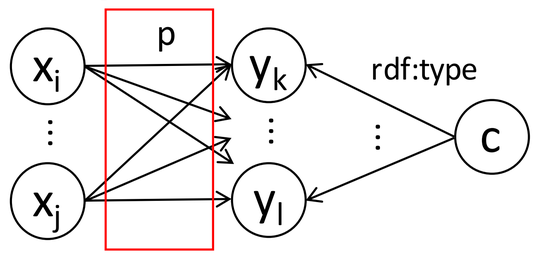
Concerns: feature spaces for classes and properties & clustering algorithm

Feature Space: Class

Internal Property Representation

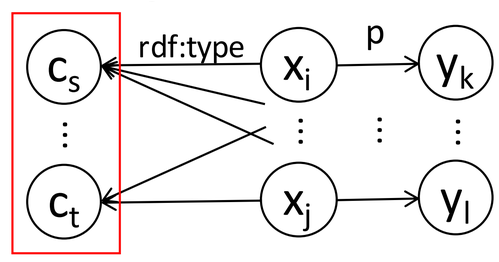


External Property Representation

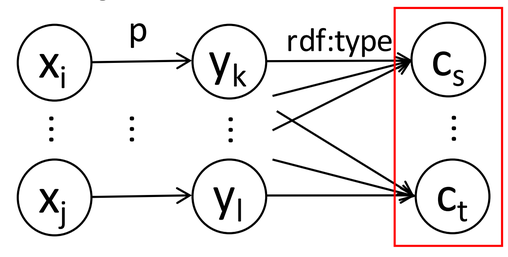


Feature Space: Property

Source Class Representation



Destination Class Representation



Algorithm

Algorithm 1 CPClustering algorithm.

Input: Classes $C^{(0)}$, Properties $P^{(0)}$
Output: Clusterings $C^{(*)}$, $P^{(*)}$

1: $t \leftarrow 0$
2: **while** $(C^{(t-1)} \neq C^{(t)}$ and $P^{(t-1)} \neq P^{(t)})$ or $t = 0$ **do**
3: $\quad C^{(t+1)} \leftarrow clustering(C^{(t)})$
4: $\quad P^{(t)} \leftarrow update(P^{(t)}, C^{(t+1)})$
5: $\quad P^{(t+1)} \leftarrow clustering(P^{(t)})$
6: $\quad C^{(t+1)} \leftarrow update(C^{(t+1)}, P^{(t+1)})$
7: $\quad t \leftarrow t + 1$
8: **end while**
9: $C^{(*)} \leftarrow C^{(t)}$, $P^{(*)} \leftarrow P^{(t)}$

(a) Class clusterings.

	IPR & SCR	IPR & DCR	EPR & SCR	EPR & DCR
IPR & SCR	-	0.30679	0.51389	0.26819
IPR & DCR	0.30679	-	0.31785	0.25950
EPR & SCR	0.51389	0.31785	-	0.27820
EPR & DCR	0.26819	0.25950	0.27820	-

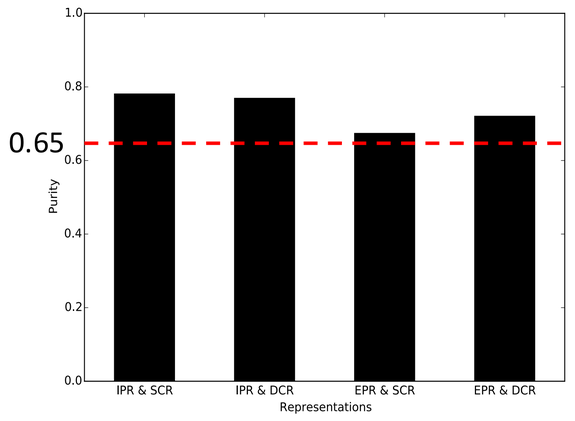
(b) Property clusterings.

	IPR & SCR	IPR & DCR	EPR & SCR	EPR & DCR
IPR & SCR	-	0.23138	0.14902	0.24907
IPR & DCR	0.23138	-	0.03130	0.81658
EPR & SCR	0.14902	0.03130	-	0.02909
EPR & DCR	0.24907	0.81658	0.02909	-

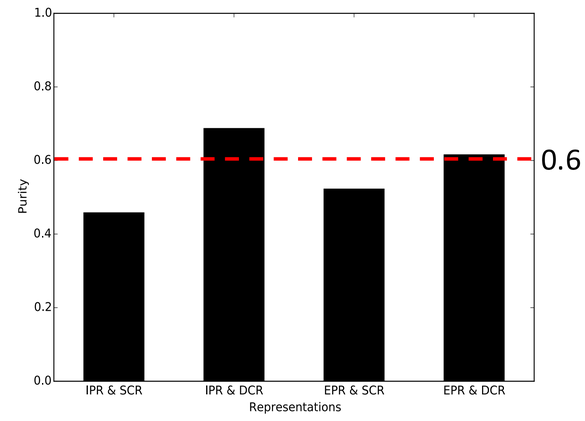
ARI among clustering w.r.t. rep.

Experimental Evaluation

- Purpose
  - Evaluate clustering effectiveness.
  - Observe differences b/w representations.
- Measurements
  - Purity (Labels are manually associated)
  - Adjusted Rand Index (ARI)
- Dataset: DBpedia



(a) Class.



(b) Property.

Purity

Future Work

- Generalize the clustering
- Revisit these representations in other aspects (e.g., probability theory)