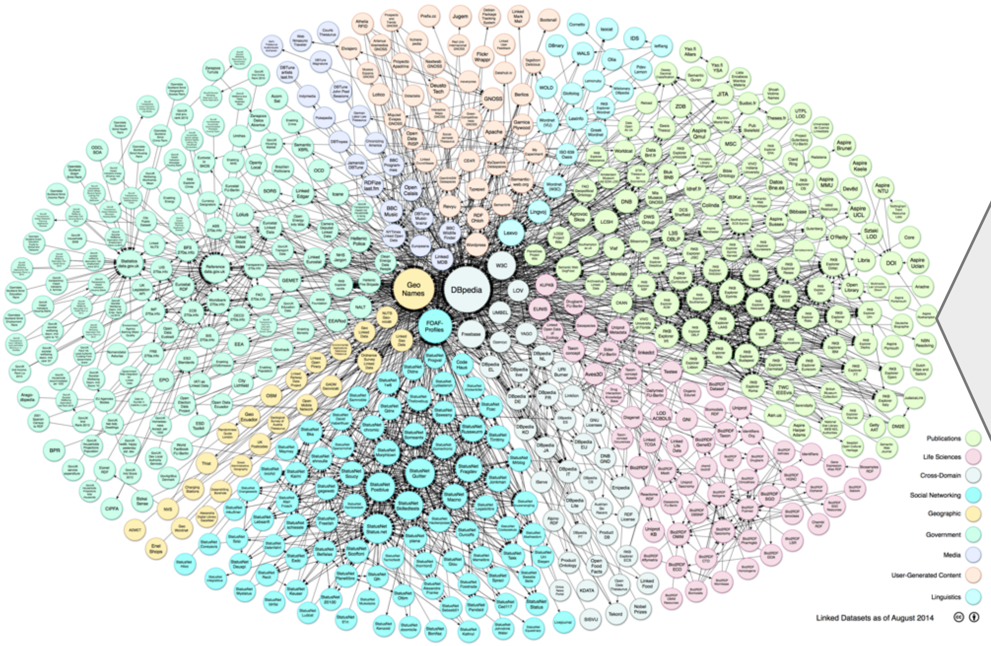
Interleaving Clustering of Classes and Properties for Disambiguating Linked Data

Takahiro Komamizu, Toshiyuki Amagasa, Hiroyuki Kitagawa

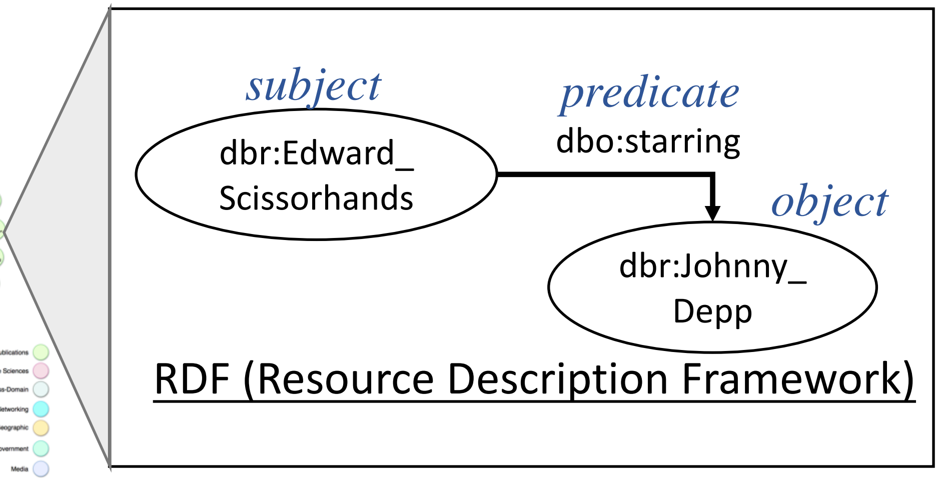
University of Tsukuba

Linked Data

- Linked Data (or LD, a.k.a. Web of data)
    - **Link** together
    - **Open** to public
    - Large number of datasets (more than 1,000 in 2014)



Linked Datasets as of August 2014



subject

dbr:Edward_ Scissorhands

predicate

dbo:starring

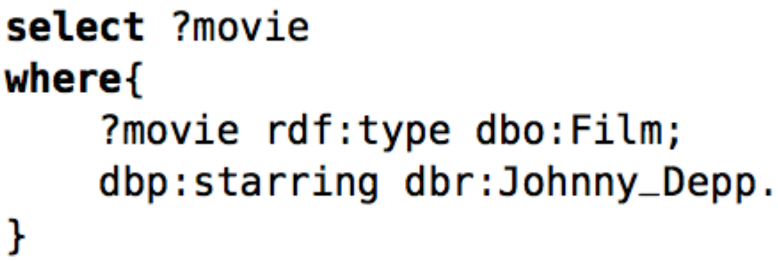
object

dbr:Johnny_ Depp

RDF (Resource Description Framework)

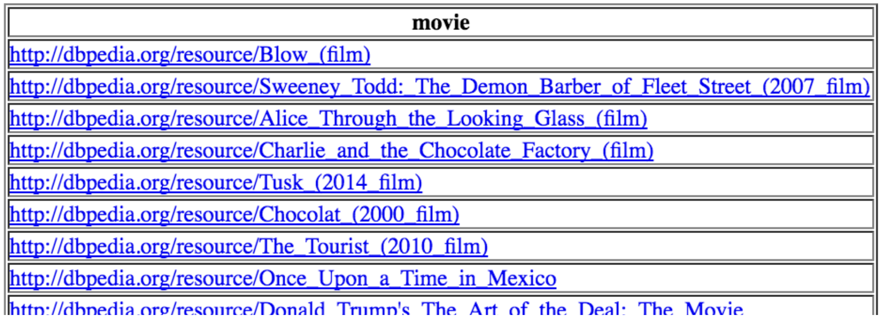
Querying via SPARQL

- SPARQL is a standardized query language for LD.



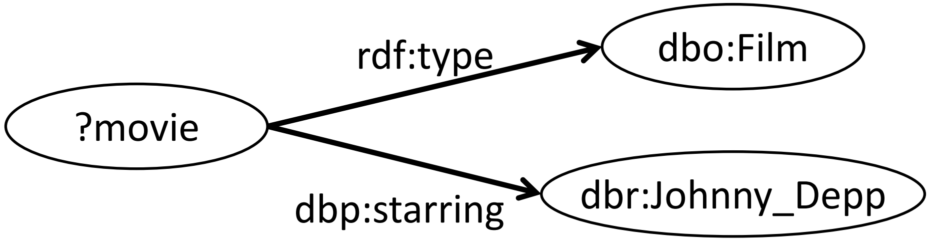
```
select ?movie
where{
    ?movie rdf:type dbo:Film;
    dbp:starring dbr:Johnny_Depp.
}
```

SPARQL



movie
http://dbpedia.org/resource/Blow_(film)
http://dbpedia.org/resource/Sweeney_Todd:_The_Demon_Barber_of_Fleet_Street_(2007_film)
http://dbpedia.org/resource/Alice_Through_the_Looking_Glass_(film)
http://dbpedia.org/resource/Charlie_and_the_Chocolate_Factory_(film)
http://dbpedia.org/resource/Tusk_(2014_film)
http://dbpedia.org/resource/Chocolat_(2000_film)
http://dbpedia.org/resource/The_Tourist_(2010_film)
http://dbpedia.org/resource/Once_Upon_a_Time_in_Mexico
http://dbpedia.org/resource/Donald_Trump's_The_Art_of_the_Deal:_The_Movie

Results



Graph representation

Ambiguities on Linked Data

- 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

These ambiguities cause

- inappropriate SPARQL queries for users
- undesired burden on adding new entities

Disambiguation with Clustering

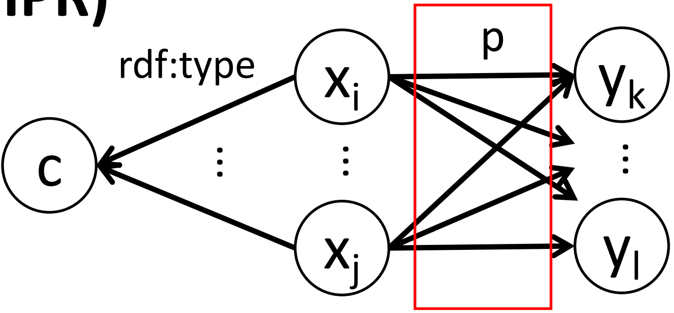
- Clustering makes groups of similar items.
- Apply clustering onto classes and properties.
    - Expectation
        - Ambiguous classes/properties compose groups.

Concerns

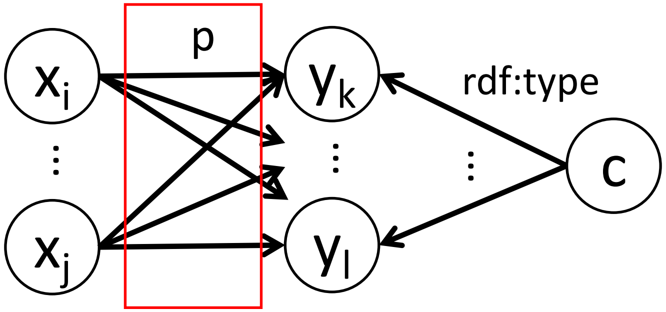
- feature spaces for classes and properties
- clustering algorithm

Feature Spaces: Class

- Classes are represented by relevant properties.
- Representations (Bag of words)
  - **Internal Property Representation (IPR)**
    - A class is represented by properties connected from instances of the class.

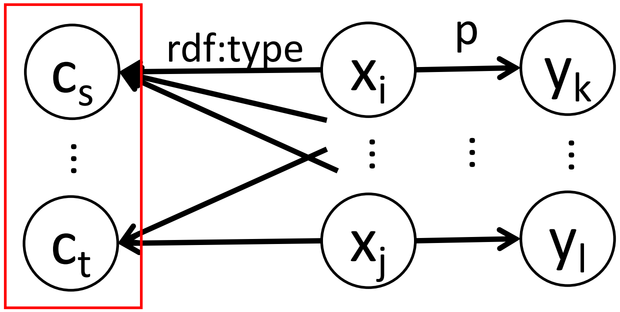


  - **External Property Representation (EPR)**
    - A class is represented by properties connecting to instances of the class.

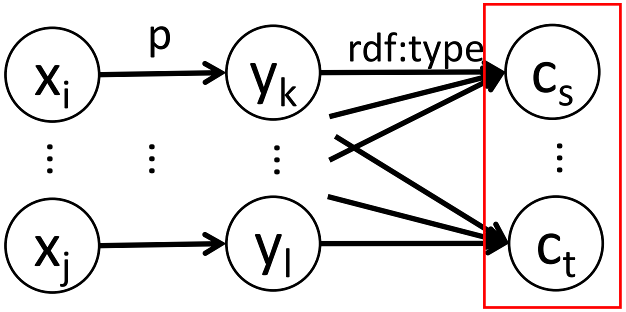


Feature Spaces: Property

- Properties are represented by relevant classes.
- Representations (Bag of words)
  - **Source Class Representation (SCR)**
    - A property is represented by classes which instances are subject of triples containing the property.

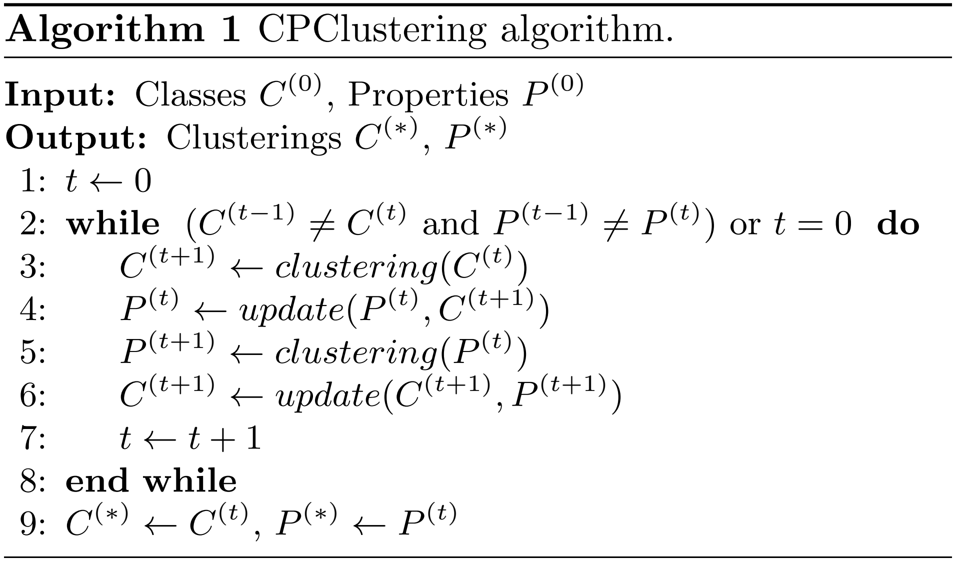


  - **Destination Class Representation (DCR)**
    - A property is represented by classes which instances are object of triples containing the property.



Interleaving Clustering: CPClustering

- As a result of the representations, clustering on classes affects properties, and vice versa.
- When classes (properties) are clustered, representations of properties (classes) are updated.



Algorithm 1 CPClustering algorithm.

```
Input: Classes $C^{(0)}$, Properties $P^{(0)}$
Output: Clusterings $C^{(*)}$, $P^{(*)}$
t ← 0
while ($C^{(t-1)} \neq C^{(t)}$ and $P^{(t-1)} \neq P^{(t)}$) or $t = 0$ do
    $C^{(t+1)} \leftarrow clustering(C^{(t)})$
    $P^{(t)} \leftarrow update(P^{(t)}, C^{(t+1)})$
    $P^{(t+1)} \leftarrow clustering(P^{(t)})$
    $C^{(t+1)} \leftarrow update(C^{(t+1)}, P^{(t+1)})$
    $t \leftarrow t + 1$
end while
$C^{(*)} \leftarrow C^{(t)}$, $P^{(*)} \leftarrow P^{(t)}$
```

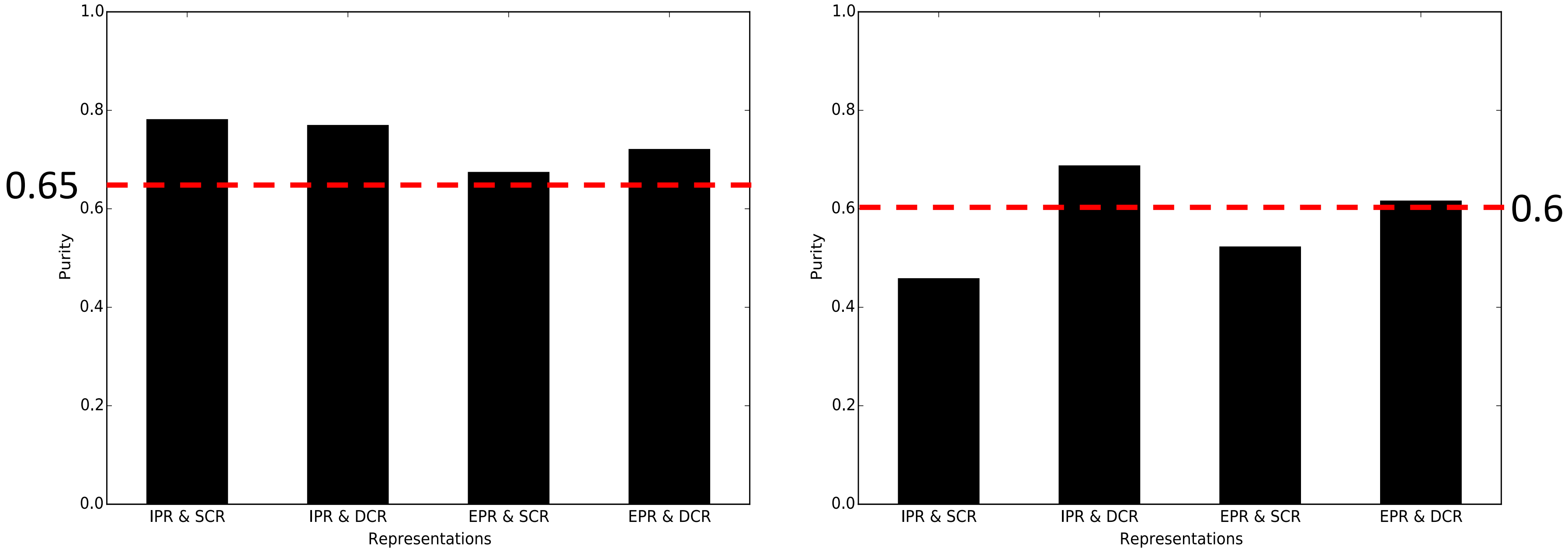
clustering

- Vector space model based similarity
  - e.g., cosine sim.
- General clustering algo.
  - e.g., k-means, DBSCAN

Experimental Evaluation

- Purpose
  - Evaluate clustering effectiveness.
  - Comparing clustering results among representations.
- Measurements
  - Purity of clusterings
    - Average on max num of same labels in each cluster
    - Labels of classes and properties are manually associated.
  - Adjusted Rand Index (ARI) between clusterings
    - ARI scores how much of item pairs are in same/different clusters.
- Dataset: classes and properties in DBpedia

Experimental Results: Purity



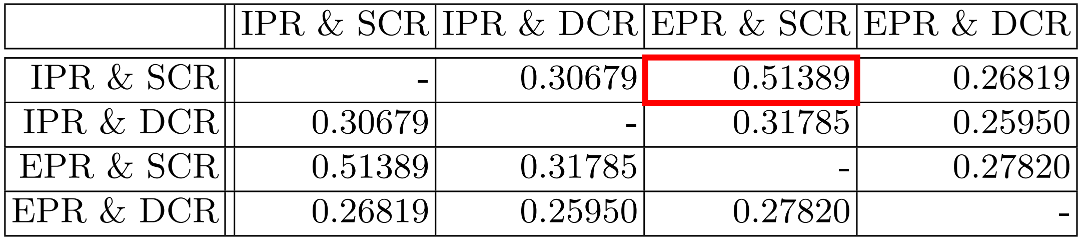
(a) Class.

(b) Property.

- Classes are well-clustered for all rep.
- Properties are well-clustered for DCR rep.

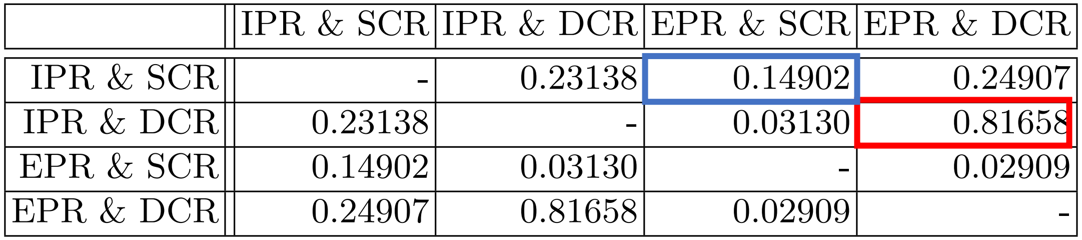
Experimental Results: ARI

(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	-

- Clusters of classes are not much overlapping.
- Clusters of properties are overlapping when DCR rep., while not overlapping when SCR rep.

➡ Still space left for improving clustering by combining these rep.

Conclusion and Future work

- CPClustering
  - Interleaving clustering of classes and properties
  - Classes (resp. properties) are represented by properties (resp. classes) in two points of views: IPR and EPR (resp. SCR and DCR)
  - Evaluation introduces reasonable purity and possibilities for combining these representations.
- Future work
  - Generalize the clustering
  - Revisit these representations in other aspects (e.g., probability theory)

Thank you for
your kind attentions.