Universität
Münster

# RDF-driven Entity Clustering of Unstructured Data

Nikolay Kazanliev

*10. July 2024*
Bachelor Thesis Presentation

# Preliminaries

## Definition (Entity)

An *entity* is a distinguishable object or concept that can be mapped to a unique identifier $i \in I$.

## Example

Example sentence: *LA is located on the West Coast*.
Found entities: $\texttt{Los\_Angeles} \in I$ and $\texttt{West\_Coast\_of\_the\_United\_States} \in I$.

# Preliminaries

## Definition (RDF triple)

An *RDF triple* is a tuple $(subject, predicate, object) \in I \times P \times (I \cup \{\mathtt{b}\})$.

## Example

$(\mathtt{Los\_Angeles}, \mathtt{part\_of}, \mathtt{West\_Coast\_of\_the\_United\_States}) \in I \times P \times I$

Universität Münster

# Preliminaries

## Definition (RDF graph)

An RDF graph $G$ is a set of RDF triples.

## Example

Consider the following graph: $G_0$:

$$
\begin{aligned}
G_0 = \{&(\texttt{Barack\_Obama}, \texttt{leader\_of\_political\_party}, \texttt{b}),\\
&(\texttt{Mahatma\_Gandhi}, \texttt{leader\_of\_political\_party}, \texttt{b}),\\
&(\texttt{Los\_Angeles}, \texttt{part\_of}, \texttt{b}),\\
&(\texttt{Los\_Angeles}, \texttt{postal\_code}, \texttt{b}),\\
&(\texttt{Berlin}, \texttt{capital\_of}, \texttt{Germany}),\\
&(\texttt{Berlin}, \texttt{postal\_code}, \texttt{b})\}
\end{aligned}
$$

# Preliminaries

## Definition (Candidate description)

Given an entity $i \in I$ and an RDF graph $G$, a *candidate description (CD)* of the entity $i$ is the set

$$CD = \{p \mid \exists o \in I \cup \{b\} : (i, p, o \in G)\}$$

## Example

$G_0$ corresponds to $\mathcal{CD}_0 = \{CD_1, CD_2, CD_3, CD_4\}$ with:

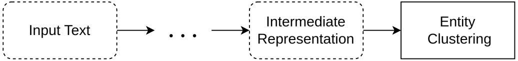
$$CD_1 = \{\texttt{leader\_of\_political\_party}\}$$
$$CD_2 = \{\texttt{leader\_of\_political\_party}\}$$
$$CD_3 = \{\texttt{part\_of}, \texttt{postal\_code}\}$$
$$CD_4 = \{\texttt{capital\_of}, \texttt{postal\_code}\}$$

Universität Münster

Introduction

# Related Work

- Text-based Approaches
  - Word embedding models (Alsudais and Tchalian, 2019)
  - Hybrid approach, including cooccurrence in a set of documents, numeric features, entity types and crowdsourcing (Lee et al., 2013)




- Graph-based Approaches
  - Hierarchical clustering on RDF datasets (Christodoulou et al., 2015; Eddamiri et al., 2019)

Universität Münster

## Introduction

# Objective

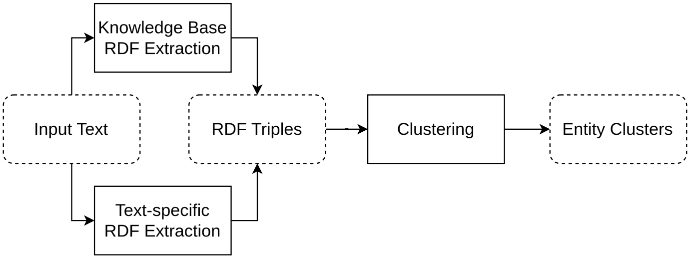



**Figure 2.1:** Initial Pipeline Structure.

Method

# Entity Clustering Algorithm

**Algorithm 3.1** Entity Clustering (Part 1). (Christodoulou et al., 2015; Eddamiri et al., 2019)

**Require:** RDF graph $G$, $cd\text{-}sim$ (similarity measure), $linkage$ (linkage method)
$cluster\text{-}score$ (clustering evaluation score)

Extract set of candidate descriptions $\mathcal{CD} = \{CD_1, \ldots, CD_{|\mathcal{CD}|}\}$ from $G$

$m \leftarrow 0$

$U^m \leftarrow \{\{CD_1\}, \{CD_2\}, \ldots, \{CD_{|\mathcal{CD}|}\}\}$

Build similarity matrix $M^m = |\mathcal{CD}| \times |\mathcal{CD}|$:

$\quad M^m_{ij} = cd\text{-}sim(CD_i, CD_j)$

Convert the similarity matrix to a distance matrix:

$\quad M^m_{ij} = 1 - M^m_{ij}$

Method

# Entity Clustering Algorithm

**Algorithm 3.1** Entity Clustering (Part 2).

**while** $m \leq |\mathcal{CD}| - 1$ **do**
  Let $(U_i^m, U_j^m)$ be the most similar pair in $U^m$ for $i \neq j$:
    $\operatorname*{argmin}_{(U_i^m, U_j^m) \in U^m} M_{ij}^m$
  $m \leftarrow m + 1$
  $U_l^m \leftarrow U_i^{m-1} \cup U_j^{m-1}$
  Update distance matrix:
    $M_{lk}^m = linkage(U_i^{m-1} \cup U_j^{m-1}, U_k^{m-1})$ for all $k \neq i, j$
  $U^m \leftarrow U^{m-1} \setminus \{U_i^{m-1}, U_j^{m-1}\} \cup U_l^m$
  $C \leftarrow C \cup U^m$

Universität Münster

## Method

# Entity Clustering Algorithm

**Algorithm 3.1** Entity Clustering (Part 3).

Let $U^i$ be the clustering with the best score:

$$\underset{U^i \in C}{\text{argmax}}\ cluster\text{-}score(U^i)$$

Map each $CD$ in $U^i$ to its unique identifier $i \in E \subseteq I$

**return** mapped $U^i$

Method

# Similarity Measures

- Jaccard Similarity

$$Jaccard(CD_i, CD_j) = \frac{|CD_i \cap CD_j|}{|CD_i \cup CD_j|} \in [0, 1]$$

- Sorensen Similarity

$$Sorensen(CD_i, CD_j) = \frac{2|CD_i \cap CD_j|}{|CD_i| + |CD_j|} \in [0, 1]$$

- Cosine Similarity

$$cosine(CD_i, CD_j) = \frac{CD_i \cdot CD_j}{\|CD_i\|\|CD_j\|} \in [0, 1]$$

# Linkage Methods for Predefined Distances

- Average Linkage

$$linkage(i \cup j, k) = \frac{n_i M_{ik} + n_j M_{jk}}{n_i + n_j}$$

- Complete Linkage

$$linkage(i \cup j, k) = \max(M_{ik}, M_{jk})$$

- Single Linkage

$$linkage(i \cup j, k) = \min(M_{ik}, M_{jk})$$

- Weighted Linkage

$$linkage(i \cup j, k) = 0.5(M_{ik} + M_{jk})$$

# Linkage Methods for Euclidean Distances

- Centroid Linkage

$$cluster\text{-}dist(i,j) = ||c_i - c_j||$$

- Median Linkage

$$cluster\text{-}dist(i,j) = ||w_i - w_j||$$
$$w_l = \frac{1}{2}(w_i + w_j)$$

- Ward Linkage

$$cluster\text{-}dist(i,j) = \sqrt{\frac{2n_i n_j}{n_i + n_j}}||c_i - c_j||$$

Method

# Clustering Evaluation Score: Silhouette Coefficient

$$Sil(CD_i) = \frac{b(CD_i) - a(CD_i)}{\max\{a(CD_i), b(CD_i)\}} \in [-1, 1]$$

where

$a(CD_i)$ is the average distance between $CD_i$ and each other element in its assigned cluster

$b(CD_i)$ is the average distance between $CD_i$ and each other element in its *neighbor* cluster

$$cluster\text{-}score(U^l) = ASW(U^l) = \sum_{i=1}^{n} \frac{Sil(CD_i)}{n}$$

## Implementation

# Overview

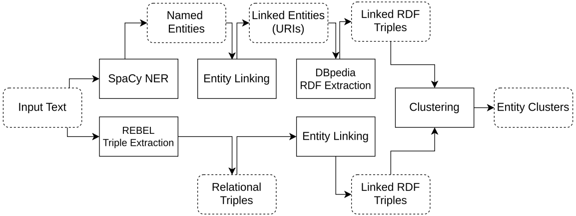



**Figure 4.1:** Pipeline Overview.

# RDF Extraction using NLP (REBEL)

## Example (Input)

In 1984, Paul Leduc released the biopic *Frida*, *naturaleza viva*, starring Ofelia Medina as Kahlo.

## Relational Triples Recognized by REBEL (Cabot and Navigli, 2021)

```
<Frida, naturaleza viva> <publication date> <1984> .
<Frida, naturaleza viva> <cast member> <Ofelia Medina> .
```

## Mapping to DBpedia URIs:

```
<http://dbpedia.org/resource/Frida_Still_Life> <publication date> <1984> .
                                          <cast member> <Ofelia Medina> .
```

# RDF Extraction from DBpedia

## RDF Triple

```
@prefix dbr: <http://dbpedia.org/resource/> .
dbr:Frida_Still_Life <http://dbpedia.org/property/producer> _:b0 .
```

## `rdf:type` value as predicate

```
dbr:Frida_Still_Life <http://schema.org/CreativeWork> _:b0 .
```

## `dcterms:subject` value as predicate

```
dbr:Frida_Still_Life dbr:Category:Biographical_films_about_painters _:b0 .
```

# Clustering

Clustering using Algorithm 3.1 and the following hyperparameters:

- Similarity measures (Section 3.2)
- Linkage methods (Section 3.3 and 3.4)
- Silhouette coefficient as a clustering evaluation metric (Section 3.5)

This results in $3 \times 7 \times 1 = 21$ clusterings

Implementation

# Labeling

Using the `gold:hypernym` DBpedia property for cluster labeling.

**Example (Labeled cluster)**

```
"Film 1": {
    "Viva_la_Vida": "Song",                     # Song
    "Broken_Wings_(Mr._Mister_song)": "Song",   # Song
    "Frida": "Film",                            # Film
    "Volver": "Not found",                      # Film
    "Frida_Still_Life": "Film",                 # Film
    "La_Flor": "Not found"                      # Film
}
```

# Manual Evaluation: Example

A cluster can be evaluated as *accurate*, *partly accurate* or *inaccurate* (Alsudais and Tchalian, 2019)

## Example (Partly Accurate Cluster)

```
"Film 1": {
    "Viva_la_Vida": "Song",                    # Song -> 0 (inacc.)
    "Broken_Wings_(Mr._Mister_song)": "Song",  # Song -> 0
    "Frida": "Film",                           # Film -> 1 (accurate)
    "Volver": "Not found",                     # Film -> 1
    "Frida_Still_Life": "Film",                # Film -> 1
    "La_Flor": "Not found" }                   # Film -> 1
```

Note: In a broader context the cluster can be evaluated as *accurate* because all elements are types of creative work.

# One-element Clusters

The relative number of one-element cluster depends on the selected linkage method.

## Example (Ward linkage)

| | Obama | Gandhi | LA | Berlin |
|---|---|---|---|---|
| Obama | 0 | **0.68** | 0.75 | 0.8 |
| Gandhi | **0.68** | 0 | 0.8 | 0.8 |
| LA | 0.75 | 0.8 | 0 | 0.7 |
| Berlin | 0.8 | 0.8 | 0.7 | 0 |

$$Sil(U^0) := ASW(U^0) = 0$$

| | Obama, Gandhi | LA | Berlin |
|---|---|---|---|
| Obama, Gandhi | 0 | 0.805 | 0.836 |
| LA | 0.805 | 0 | **0.7** |
| Berlin | 0.836 | **0.7** | 0 |

$$Sil(U^1) = 0.0608$$

| | Obama, Gandhi | LA, Berlin |
|---|---|---|
| Obama, Gandhi | 0 | 0.875 |
| LA, Berlin | 0.875 | 0 |

$$Sil(U^2) = \mathbf{0.1236}$$

# One-element Clusters

## Example (Median linkage)

| | Obama | Gandhi | LA | Berlin |
|---|---|---|---|---|
| Obama | 0 | **0.68** | 0.75 | 0.8 |
| Gandhi | **0.68** | 0 | 0.8 | 0.8 |
| LA | 0.75 | 0.8 | 0 | 0.7 |
| Berlin | 0.8 | 0.8 | 0.7 | 0 |

$$Sil(U^0) := ASW(U^0) = 0$$

| | Obama, Gandhi | LA | Berlin |
|---|---|---|---|
| Obama, Gandhi | 0 | **0.697** | 0.724 |
| LA | **0.697** | 0 | 0.7 |
| Berlin | 0.724 | 0.7 | 0 |

$$Sil(U^1) = \mathbf{0.0608}$$

| | Obama, Gandhi, LA | Berlin |
|---|---|---|
| Obama, Gandhi, LA | 0 | 0.621 |
| Berlin | 0.621 | 0 |

$$Sil(U^2) = 0.0211$$

$$Sil(CD_{\mathtt{LA}}) = -0.0968$$

# Evaluation Metrics

For a clustering $U^k$ with $n = |U^k|$ and $U_i \in U^k$:

- Coherence Measure

$$\text{CM}_{cluster}(U_i) = \frac{\text{Number of relevant entities in } U_i}{\text{Number of entities in } U_i} \in [0, 1]$$

$$\text{CM}_{overall}(U^k) = \frac{1}{n} \sum_{j=1}^{n} CM_{cluster}(U_j) \in [0, 1]$$

- Precision Measure

$$\text{PM}(U^k) = \frac{\sum_{j=1}^{n} \text{Number of relevant entities in } U_j}{\text{Total number of entities in } U^k} \in [0, 1]$$

# Evaluation Metrics

- Coherent Clusters Measure

$$\text{CCM}(U^k) = \frac{\text{Number of (partly) accurate clusters in } U^k}{n} \in [0, 1]$$

- Weighted Precision Measure

$$\text{WPM}(U^k) = \text{PM}(U^k) \left( \frac{n - \text{ Number of one-element clusters in } U^k}{n} \right)$$

WPM is used as a counterbalance for CM, PM, CCM (Alsudais and Tchalian, 2019).

Evaluation

# Evaluation Results

Results for the Germany Wikipedia article[1] as a representative example:

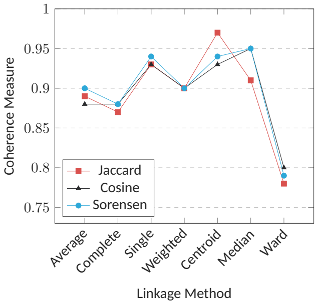



**Figure 5.1:** Coherence Measure.

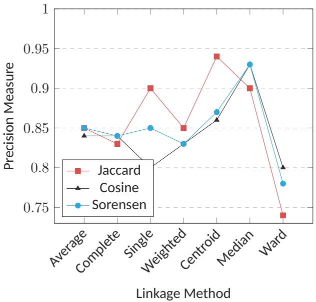



**Figure 5.2:** Precision Measure.

[1] https://en.wikipedia.org/wiki/Germany

Evaluation

# Evaluation Results

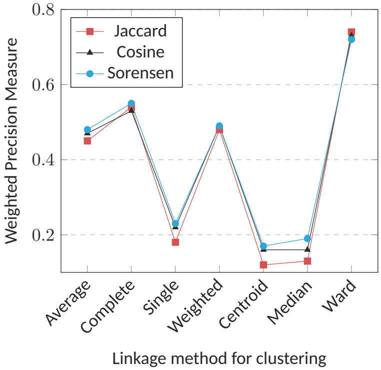



**Figure 5.3:** Weighted Precision Measure.

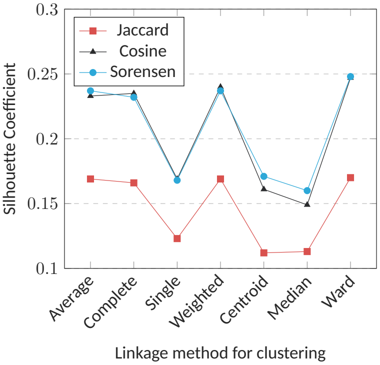



**Figure 5.4:** Silhouette Coefficient.

# Evaluation: Remarks

- Consistent results across all similarity measures
- Silhouette coefficient can serve as a predictor for Weighted Precision Measure.
- In combination with the Silhouette coefficient:
  - Median, centroid and single linkage achieve high accuracy but result in fewer clustered entities.
  - Average, complete, weighted and Ward linkage generate better-defined clusterings.

Universität Münster

# Discussion

- Text-specific relations have a minor influence on the formed clusters due to the number of DBpedia triples being much larger than the number of triples extracted through NLP.
- Error propagation

**Example (Error propagation)**

`<Bobby Rush> <member of political party> <Democratic> .`

The subject of the relational triple is linked to a false URI:

```
<http://dbpedia.org/resource/Rush_(band)>
    <member of political party> <Democratic> .
```

This error results in unwanted clustering and labeling:

```
"Band 1": {
        "United_States_Armed_Forces": "Forces",
        "Rush_(band)": "Band"}
```

# Conclusion

- RDF data can serve as an effective intermediate semantic representation for clustering.
- Future work areas include:
  - Replacing REBEL with an NLP relation extraction (RE) model that has a higher extraction rate for text-specific relations.
  - Assigning higher weight to NLP-extracted triples given a highly accurate RE model.
  - Incorporating other open knowledge bases, such as YAGO or Wikidata.
  - Restricting predicates from open knowledge bases to a predefined subset of relevant ones.

# Thank you for your attention!

# References

Abdulkareem Alsudais and Hovig Tchalian. Clustering prominent named entities in topic-specific text corpora. In *25th Americas Conference on Information Systems, AMCIS 2019, Cancún, Mexico, August 15-17, 2019*. Association for Information Systems, 2019.

Pere-Lluís Huguet Cabot and Roberto Navigli. REBEL: relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2370–2381. Association for Computational Linguistics, 2021. DOI: 10.18653/V1/2021.FINDINGS-EMNLP.204.

Klitos Christodoulou, Norman W. Paton, and Alvaro A. A. Fernandes. Structure inference for linked data sources using clustering. *Trans. Large Scale Data Knowl. Centered Syst.*, 19:1–25, 2015. DOI: 10.1007/978-3-662-46562-2_1.

Siham Eddamiri, El Moukhtar Zemmouri, and Asmaa Benghabrit. An improved rdf data clustering algorithm. In *Procedia Computer Science*, volume 148, 2019. DOI: 10.1016/j.procs.2019.01.038.

Jongwuk Lee, Hyunsouk Cho, Jin-Woo Park, Young-rok Cha, Seung-won Hwang, Zaiqing Nie, and Ji-Rong Wen. Hybrid entity clustering using crowds and data. *VLDB J.*, 22(5):711–726, 2013. DOI: 10.1007/S00778-013-0328-8.