Web Of Data: Exam 2024

Pablo Mollá Chárlez

Contents

1 Part 1: Data Linking (9pts)

- **Question 1 (2 pts).** Give three main families of data linking approaches and, for each, give its main characteristics.
- **Question 2. (2 pts)** What are the main aspects that may be considered for evaluating data linking approaches?

  Let us consider two datasets D_1 and D_2 shown in pictures 2 and **??** which give an extract of some film descriptions. These films are described by five properties $\{\text{title}, \text{hasActor}, \text{rDate}, \text{director}, \text{lang}\}$. We note that the properties hasActor* and director* are multi-valued and we consider that for each pair of equal values we have a starting **synVals**:

$$\text{synVals}(\text{"Ocean's 11"}, \text{"Ocean's 11"}), \text{synVals}(\text{"2004"}, \text{"2004"}),$$
$$\text{synVals}(\text{"P. Greengrass"}, \text{"P. Greengrass"}), \text{synVals}(\text{"J. Roberts"}, \text{"J. Roberts"}),$$

	title	hasActor*	rDate	director*	lang
i_1	Ocean's 11	J. Roberts; B. Pitt;	2001	S. Soderbergh	
i_2	Ocean's 12	J. Roberts; B. Pitt; G. Clooney	2004	S. Soderbergh; P.Greengrass	
i_3	Ocean's 13	B. Pitt; G. Clooney	2007	S. Soderbergh	
i_4	The descendants	N. Krause; G. Clooney	2011	A. Payne	en
i_5	Bourne Identity		2002	P. Greengrass	en
i_6	Ocean's twelve	J. Roberts; B. Pitt; G. Clooney	2004		

Figure 1: Extract of film descriptions dataset D_1

	title	hasActor	rDate	director	lang
i_{12}	Ocean's 11	J. Roberts; B. Pitt;	2001	S. Soderbergh	
i_{22}	Ocean's 12	J. Roberts; B. Pitt	2004	S. Soderbergh; P.Greengrass	
i_{32}	Ocean's 13	B. Pitt; G. Clooney	2007	S. Soderbergh; P.Greengrass	
i_{52}	Bourne Identity		2002	P. Greengrass	en
i_{62}	Ocean's twelve	J. Roberts; B. Pitt; G. Clooney	2004		

Figure 2: Extract of film descriptions dataset D_2

- **Question 3 (3 pts).** Using the L2R method and considering the axiom PFI(hasActor, director) of the class Film what would be the **owl:sameAs** links that can be obtained between the instances of D_1 and D_2?
- **Question 4 (2 pts).** If you apply the property sharing rule of the **sameAs** predicate:

$$sameAs(x, y) \wedge p(x, z) \rightarrow p(y, z)$$

  What would be the new property values that can be inferred?

1.1 Answers

- **Question 1 (2 pts).** Give three main families of data linking approaches and, for each, give its main characteristics.

  The (four) families of data linking approaches with its characteristics are:

  - **Instance-Based Approaches:** It focuses solely on data type properties also known as attributes and utilises similarity measures such as Jaccard or Levenshtein distance, to match entites based on their attribute values. Examples of tools for such approach include SILK or KNOFUSS. The first, provides a link specification language (LSK) for specifying linking rules using similarity measures and thresholds. The second implements unsupervised attribute-based similarity measures.

- **Graph-Based Approaches:** This approach considers both data type properties (attributes) and object properties (relations). It propagates similarity scores or linking decisions through realationships in the graph enabling collective data linking. Relies on ontology axioms to guide and refine linking. The tools of such approach include L2NR framework which combines logical and numerical rules methods for reconciliation like disjunctions and functionality to filter or infer links.
- **Supervised Approaches:** Requires labeled data of expert-defined samples of linked entities. It uses the samples to train machine learning linking rules and involves manual or interactive input ti build reliable training sets but requires significant human effort to prepare.
- **Rule-Based Approaches:** Relies on explicit knowledge encoded in ontologies or provided by domain experts. Requieres specification of rules and can work well when domain knowledge is abundant but struggles with scalability and adaptability to new datasets.

- **Question 2. (2 pts)** What are the main aspects that may be considered for evaluating data linking approaches?

  The main aspects when evaluating data linking approaches include:

  - **Effectiveness:** This measures the quality of the linking results, often quantified using:
    * Recall: Recall = $\frac{\text{\#correct-links-sys}}{\text{\#correct-links-groundtruth}}$, which represents the proportion of correct links identified by the system out of the total correct links in the ground truth.
    * Precision: Precision = $\frac{\text{\#correct-links-sys}}{\text{\#links-sys}}$, which represents the proportion of correct links identified by the system out of all the links it predicted.
    * F-measure (F1): $F1 = \frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall}+\text{Precision}}$, which provides a harmonic mean of precision and recall to balance their trade-offs.
  - **Efficiency:** Refers to the computational performance of the approach in terms of time and space. The goal is to minimize the linking search space and the interaction required with experts or users.
  - **Robustness:** Measures the system's ability to handle and override errors or inconsistencies in the input data. A robust system should still perform well despite noisy or incomplete data.
  - **Use of Benchmarks:** The evaluation may rely on standardized benchmarks, such as those provided by the Ontology Alignment Evaluation Initiative (OAEI) or Lance, to ensure consistent and objective performance comparisons across systems. These benchmarks include datasets and tasks specifically designed to test the effectiveness and robustness of data linking approaches.

- **Question 3 (3 pts).** Using the L2R method and considering the axiom PFI(hasActor, director) of the class Film what would be the owl:sameAs links that can be obtained between the instances of D_1 and D_2?

  From the information given about both datasets D_1 and D_2, and mostly handling any pair of equal values as synVals, then we have the following results using the provided axiom:

$$\textbf{PFI(hasActor, director)} \longleftrightarrow \text{synVals}(X_1, X_2) \wedge \text{synVals}(Y_1, Y_2) \wedge \text{hasActor}(X_1, X) \wedge \text{hasActor}(X_2, Y) \wedge \text{director}(Y_1, X) \wedge \text{director}(Y_2, Y) \implies \text{sameAs}(X, Y)$$

  The **sameAs** instances found are:

$$\bigstar \begin{cases} \text{synVals(”P. Greengrass”, ”P. Greengrass”)} \wedge \text{synVals(”J. Roberts”, ”J. Roberts”)} \wedge \\ \text{hasActor}(i_2, \text{”J. Roberts”}) \wedge \text{hasActor}(i_{22}, \text{”J. Roberts”}) \wedge \text{director}(i_2, \text{”P. Greengrass”}) \wedge \\ \text{director}(i_{22}, \text{”P. Greengrass”}) \underbrace{\implies}_{PFI} \mathit{sameAs}(i_2, i_{22}) \end{cases}$$

$$\bigstar\bigstar \begin{cases} \text{synVals(”S. Soderbergh”, ”S. Soderbergh”)} \wedge \text{synVals(”J. Roberts”, ”J. Roberts”)} \wedge \\ \text{hasActor}(i_1, \text{”J. Roberts”}) \wedge \text{hasActor}(i_{12}, \text{”J. Roberts”}) \wedge \text{director}(i_1, \text{”S. Soderbergh”}) \wedge \\ \text{director}(i_{12}, \text{”S. Soderbergh”}) \underbrace{\implies}_{PFI} \mathit{sameAs}(i_1, i_{12}) \end{cases}$$

$$\bigstar\bigstar\bigstar \left\{ \begin{array}{l} \text{synVals("S. Soderbergh", "S. Soderbergh")} \ \wedge \ \text{synVals("B. Pitt", "B. Pitt")} \ \wedge \\ \text{hasActor}(i_3, \text{"B. Pitt"}) \ \wedge \ \text{hasActor}(i_{32}, \text{"B. Pitt"}) \wedge \ \text{director}(i_3, \text{"S. Soderbergh"}) \ \wedge \\ \text{director}(i_{32}, \text{"S. Soderbergh"}) \ \underbrace{\Longrightarrow}_{PFI} \ sameAs(i_3, i_{32}) \end{array} \right.$$

- **Question 4 (2 pts).** If you apply the property sharing rule of the **sameAs** predicate:

$$sameAs(x, y) \ \wedge \ p(x, z) \ \rightarrow \ p(y, z)$$

What would be the new property values that can be inferred?

By applying this new consideration, we obtain the following sameAs instances:

- $sameAs(i_2, i_{22}) \ \wedge \ hasActor(i_2, \text{"}G.\ Clooney\text{"}) \ \rightarrow \ hasActor(i_{22}, \text{"}G.\ Clooney\text{"})$
- $sameAs(i_3, i_{32}) \ \wedge \ p(i_{32}, \text{"}P.\ Greengrass\text{"}) \ \rightarrow \ p(i_3, \text{"}P.\ Greengrass\text{"})$

2 Part 2: Ontology Alignment (6pts)

- **Question 5 (1.5 pt).** Give three kinds of heterogeneity in ontologies that can be faced when dealing with ontology alignment.
- **Question 6 (2 pts).** Given the ontology alignment problem shown in Figure 3:
  1. Explain the different inputs.
  2. Give two examples of relations that can be used to represent mappings in A'.

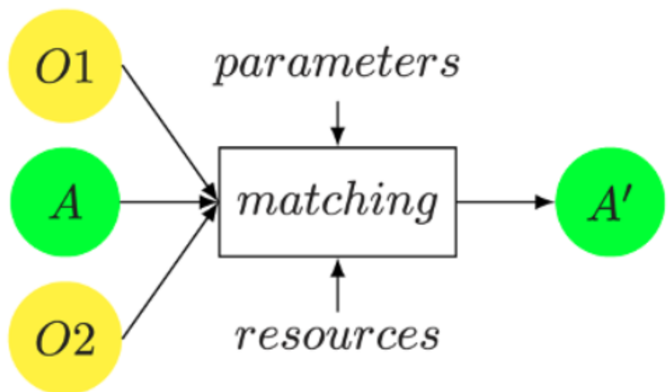

Figure 3: Ontology Alignment Problem

Let us consider two ontologies O_1 and O_2 of Figure 4. In table 5, we give the set of identity links between instances of these two ontologies.

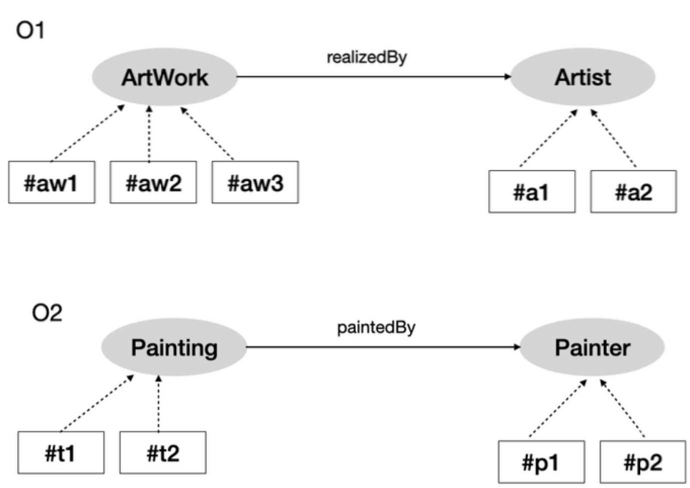

Figure 4: Ontologies O_1 and O_2

SameAs(#aw1, #t1)	SameAs(#aw2, #t2)
SameAs(#a1, #p1)	SameAs(#a2, #p2)

Figure 5: Identity Links of the Instances of O_1 and O_2

- **Question 7 (1.5 pt).** If we apply an instance-based ontology alignment what would be the ontology mappings between the classes of these two ontology that can be found?

- **Question 8 (1 pt).** In the same setting, what would be the ontology mappings between the properties of these two ontologies that you may suggest?

2.1 Answers

- **Question 5 (1.5 pt).** Give three kinds of heterogeneity in ontologies that can be faced when dealing with ontology alignment.

  The 3 kinds of heterogeneity in ontologies that can be faced when dealing with ontology alignment are:

  1. **Syntactic Heterogeneity:** The differences in the format of representation of ontologies such as varying description languages (RDF, RDFS, OWL).
  2. **Terminological Heterogeneity:** This includes the variations in labels, names and terminology used to describe the same concepts or entities (e.g. "car" and "automobile").
  3. **Conceptual Heterogeneity:** Finally, this heterogeneity considers the differences in the coverage, granularity or perspective of the modeled concepts.

- **Question 6 (2 pts).** Given the ontology alignment problem shown in Figure 3:

  1. (1) Explain the different inputs.
  2. (2) Give two examples of relations that can be used to represent mappings in A'.

  (1) First of all, let's define the different inputs involved in the image. O_1 and O_2 are the input ontologies that need to be aligned, they represent structured knowledge in specific domains. A represents an initial set of mappings between both ontologies, which might be incomplete or approximate. It acts as a starting point to refine or extend the matching process. Parameters are the criteria or thresholds guiding the matching process like similarity metrics or alignment constraints. Resources represent the auxiliary resources used to enhance the matching process such as external vocabularies, background knowledge or machine learning models. Finally, A' is the result of the matching process which includes mappings specifying how entities from O_1 and O_2 are related.

  (2) Secondly, let's mention some types of relations that can connect ontologies via mappings.

  1. **Equivalent Class:** It specifies that a class in O_1 is semantically equivalent to a class in O_2 (**equivalentClass**).
  2. **Equivalent Property:** It specifies that a property in O_1 corresponds to a property in O_2 with the same meaning (**equivalentProperty**).
  3. **Subsumption:** It indicates a hierarchical relationship where a class/property in O_1 is a subclass/subproperty of one in O_2 (**rdfs:subClassOf** or **rdfs:subPropertyOf**)
  4. **Instance Matching:** It denotes that 2 instances refer to the same entity.

- **Question 7 (1.5 pt).** If we apply an instance-based ontology alignment what would be the ontology mappings between the classes of these two ontology that can be found?

  By applying an instance-based ontology alignment, we would obtain the following sameAs links:

  O_2**:Painting** **SubClassOf** O_1**:ArtWork**

  As well as, the following equivalent classes, because there is a bijection between the sets of both instances classes:

  O_2**:Painter** **equivalentClass** O_1**:Artist**

- **Question 8 (1 pt).** In the same setting, what would be the ontology mappings between the properties of these two ontologies that you may suggest?

  I would personally suggest the following mapping:

  $$O_2\textbf{:paintedBy} \quad \textbf{subPropertyOf} \quad O_1\textbf{:realizedBy}$$

3 Part 3: Link Validation (5pts)

- Question 9 (1.5 pts). Give three reasons that may lead to incorrect **sameAs** links.
- **Question 10 (1.5 pts).** Give the four properties that define the semantics of **sameAs** predicate.
- **Question 11 (2 pts).** According to the recent literature studies, cite three different kinds of approaches that can be used to detect erroneous identity links.

3.1 Answers

- Question 9 (1.5 pts). Give three reasons that may lead to incorrect sameAs links.

  The following reasons might lead to incorrect sameAs links:

  1. **Resource or Terminological Ambiguities:** Different entities may have a similar or identical label, leading to false matches such as "Paris" considered as the city and "Paris" considered as the mythological figure.
  2. **Structural or Ontological Differences:** Variations in the modelling of concepts across datasets or ontologies can create mismatches.
  3. **Data Source Quality & Trustworthiness:** Errors in the original data or low-quality sources may propagate incorrect identity assertions.

- **Question 10 (1.5 pts).** Give the four properties that define the semantics of **sameAs** predicate.

  The definition of the sameAs predicate is based on:

  1. **Reflexivity:** Every resource is the same as itself. $\forall X \; owl: sameAs(X, X)$
  2. **Symmetry:** If a resource X is the same Y, then Y is also the same as X. $\forall X, Y \; owl: sameAs(X, Y) \implies owl: sameAs(Y, X)$
  3. **Transitivity:** $\forall X, Y, Z \; owl: sameAs(X, Y) \wedge owl: sameAs(Y, Z) \implies owl: sameAs(X, Z)$
  4. **Property Sharing:** If X is the same as Y and X has a property Z, then Y must also have Z. $\forall X, Y, Z \; owl: sameAs(X, Y) \wedge P(X, Z) \implies P(Y, Z)$

- **Question 11 (2 pts).** According to the recent literature studies, cite three different kinds of approaches that can be used to detect erroneous identity links.

  Finally, we can distinguish 3 different kinds of approaches to detect erroneous identity links:

  1. **Inconsistency-Based Approaches:** These rely on detecting violations of logical assumptions or axioms such as **UNA** or ontology axioms like functional properties or inverse functional properties.
  2. **Context-Based Approaches:** These analyze the content or features of linked resources such as types, properties and textual values to identify patterns or outliers.
  3. **Network-Based Approaches:** These leverage the connectivity and metrics of nodes in the identity network to assess the quality of links.