Exam KDGD 2023/2024

Pablo Mollá Chárlez

February 15, 2025

Contents

1 Part 1: Key Discovery for Data Linking [12 pts]

1.1 Question 1 [2.5 pts]

Give three different quality measures that are used to evaluate the quality of discovered keys that are proposed by [Soru et al. 15] and [Atencia et al. 12]. For each measure, give an informal definition.

1.1.1 Answer

The three measures commonly used to assess discovered keys are support, discriminability, and score. In the context of **F-Keys** as proposed by Soru et al. (2015), and **SF-Keys** (or Preside-Keys) as discussed by Atencia et al. (2012), these measures can be described informally as follows:

- Support: This measure tells you how widely the key applies by computing the fraction of instances in the dataset that actually provide values for the key's properties. In other words, support gauges the "coverage" of the key across the data, indicating its overall applicability. Formally, the support is defined as:

$$\text{support(P)} = \frac{\text{\# instances described by P}}{\text{\# all instances}}$$

- Discriminability: Discriminability looks at how effectively the key distinguishes individual entities. It does this by partitioning the dataset according to the key's values and then checking what proportion of these groups contain exactly one instance. A discriminability of 1 means every group is a singleton—i.e., the key perfectly differentiates every record. Formally, the discriminability is defined as:

$$\text{dis(P)} = \frac{\text{\# singleton partitions}}{\text{\# partitions}}$$

- Score: Often derived as the ratio of discriminability to the total number of instances, the score provides a normalized measure of the key's quality. A score of 1 indicates a perfect key (all instances are uniquely identified), while a score below 1 shows that some duplicates exist, meaning the key is only a pseudo-key. This measure is especially useful for comparing keys across datasets. Formally, the score is defined as:

$$\text{score(P)} = \frac{\text{discriminality(P)}}{\text{\# instances}}$$

  If $score(P) = 1$, then P is a perfect key, and if $score(P) < 1$, then P is only a pseudo-key.

These quality measures together help evaluate both how broadly a key can be applied (support) and how precisely it differentiates between entities (discriminability and score).

1.2 Question 2 [2.5 pts]

To distinguish between the three key semantics S-Keys, SF**-Keys** and F-Keys that are studied in the course,

- (a) What are the main data characteristics that should be taken into account ?
- (b) How these characteristics are considered in the S-Keys, SF**-Keys** and F-Keys semantics?

1.2.1 Answer

- (a) The main data characteristics to consider are how multi-valued properties and missing values (or incomplete descriptions) are treated. In other words, one must decide whether to compare the values for a property by looking at overlaps between sets or by requiring an exact match, and whether an empty (or missing) value should be regarded as distinct from any non-empty value or possibly identical to other empty values.

- (b) In the S-Keys semantics, multi-valued properties are compared in a relaxed way—that is, two sets are considered equal if they share at least one common element—and empty values are treated as different from any existing value (an optimistic stance). In contrast, F-Keys require an exact match between the sets of values, so two instances' multi-valued properties must be identical for them to be deemed equal; moreover, empty values are treated pessimistically, meaning they might be considered identical. Finally, SF-**Keys** also require an exact match for multi-valued properties like F-Keys but adopt an optimistic approach to empty values by treating them as different from non-empty values. This combination leads to a subtle yet important difference in how keys are discovered and verified.

1.3 Question 3 [7 pts]

In **table 1** we give an extract of some book descriptions. These books are described by six properties {**title**, **hasAuthor**, **genre**, **pages**, **publisher**, **lang**}. Given these data if we apply SAKey, a key discovery tool that allows to discover n-almost keys (under the S-key semantics):

- (a) Give a **2-almost** key of one property that can be discovered. [**1.5pts**]
- (b) Give a **3-almost** key composed of two properties. [**1.5pts**]
- (c) Give a S-Key, composed of two properties, that is not an F-Key that can be discovered in the data presented in **Table 1**. [1pts]

	title	author	genre	pages	publisher	lang
b_1	The Age of Wrath	E. Abraham, Oram Andy	history	198	Penguin	en
b_2	The Trial	Kafka Frank, J. Clarck	fiction	198	R. House	
b_3	Statistical Decision Theory	Pratt John, Tao Terence	data_science	236	MIT Press	en, de
b_4	Data Mining Handbook		data_science	242	Apress	
b_5	The New Machiavelli	Wells H. G.	fiction	198	Penguin	en
b_6	Analysis & Vol I	Tao Terence, N. Robert	science	250	Apress	en
b_7	Philosophie der Physik	Heisenberg Werner	science	197	Penguin	de
b_8	Making Software		computer_science	232	O'Reilly	
b_9	Analysis & Vol I	Tao Terence	mathematics	248	HB	en

Figure 1: Extract of Books Descriptions **(D1)**

- Let consider a key $K1 = hasKey(Book)$(**title**, **hasAuthor**) and $K2 = hasKey(Book)$(**hasAuthor**, **lang**). Let **D2** be a second dataset given in **Table 2**. What would be the sameAs links that can be inferred when applying **K1** and **K2** to $D1 \times D2$ S-Key semantics. Give separated results for each key. [**3pts**]

	title	author	genre	pages	publisher	lang
b_{21}	The Age of Wrath	Eraly Abraham	history	198	Penguin	
b_{22}	Statistical Decision Theory	Pratt John	data_science	236	MIT Press	en, de
b_{23}	The New Machiavelli		fiction	198	Penguin	en
b_{24}	Philosophie der Physik	Heisenberg Werner	science	197	Penguin	de
b_{25}	Analysis & Vol I	Tao Terence	mathematics	248	HB	

Figure 2: Book Descriptions **(D2)**

1.3.1 Answer

- (a) The 2-almost key of one property that can be discovered is **{title}**. Notice that the 2 exceptions are in the instances b_6 and b_9.
- (b) The 3-almost key composed of two properties that can be discovered is **{genre, pages}**. Notice that with those 2 properties we have an 2-almost S-Key, which qualifies obviously as a 3-almost S-Key.
- (c) A S-Key, composed of two properties, that is not an F-Key that can be discovered in the data is **{author, genre}**. Notice that, even though the **author** property has on its won 3 exceptions (b_3, b_6, b_9) when adding the property **genre**, there are no visible exceptions under S-Key semantics because of the

optimisitc approach. However, under **F-Key** semantics, as the approach is pessimistic, the empty cells are considered as possible identical matches, and therefore the instance b_4 in the property **author** could have the exact same value **{Pratt John, Tao Terence}** (as in instance b_3) which would make an exception because **genre** has already the same value for b_3 and b_4 instances.

	title	author	genre	pages	publisher	lang
b_1	The Age of Wrath	E. Abraham, Oram Andy	history	198	Penguin	en
b_2	The Trial	Kafka Frank, J. Clarck	fiction	198	R. House	
b_3	Statistical Decision Theory	Pratt John, Tao Terence	data_science	236	MIT Press	en, de
b_4	Data Mining Handbook		data_science	242	Apress	
b_5	The New Machiavelli	Wells H. G.	fiction	198	Penguin	en
b_6	Analysis & Vol I	Tao Terence, N. Robert	science	250	Apress	en
b_7	Philosophie der Physik	Heisenberg Werner	science	197	Penguin	de
b_8	Making Software		computer_science	232	O'Reilly	
b_9	Analysis & Vol I	Tao Terence	mathematics	248	HB	en

Figure 3: Solutions

- (d) The following image highlights the **sameAs** links that can be inferred using key K_1 = **hasKey(Book)(title, author)**. Notice that, there are 2 different colours for the instance b_{25} as it can be inferred the **sameAs**(b_6, b_{25}) and **sameAs**(b_9, b_{25}).

	title	author	genre	pages	publisher	lang
b_1	The Age of Wrath	E. Abraham, Oram Andy	history	198	Penguin	en
b_2	The Trial	Kafka Frank, J. Clarck	fiction	198	R. House	
b_3	Statistical Decision Theory	Pratt John, Tao Terence	data_science	236	MIT Press	en, de
b_4	Data Mining Handbook		data_science	242	Apress	
b_5	The New Machiavelli	Wells H. G.	fiction	198	Penguin	en
b_6	Analysis & Vol I	Tao Terence, N. Robert	science	250	Apress	en
b_7	Philosophie der Physik	Heisenberg Werner	science	197	Penguin	de
b_8	Making Software		computer_science	232	O'Reilly	
b_9	Analysis & Vol I	Tao Terence	mathematics	248	HB	en

Figure 4: K_1 sameAs Links (D1)

	title	author	genre	pages	publisher	lang
b_{21}	The Age of Wrath	Eraly Abraham	history	198	Penguin	
b_{22}	Statistical Decision Theory	Pratt John	data_science	236	MIT Press	en, de
b_{23}	The New Machiavelli		fiction	198	Penguin	en
b_{24}	Philosophie der Physik	Heisenberg Werner	science	197	Penguin	de
b_{25}	Analysis & Vol I	Tao Terence	mathematics	248	HB	

Figure 5: K_1 sameAs Links (D2)

The following image highlights the **sameAs** links that can be inferred using key K_2 = **hasKey(Book)(author, lang)**.

	title	author	genre	pages	publisher	lang
b_1	The Age of Wrath	E. Abraham, Oram Andy	history	198	Penguin	en
b_2	The Trial	Kafka Frank, J. Clarck	fiction	198	R. House	
b_3	Statistical Decision Theory	Pratt John, Tao Terence	data_science	236	MIT Press	en, de
b_4	Data Mining Handbook		data_science	242	Apress	
b_5	The New Machiavelli	Wells H. G.	fiction	198	Penguin	en
b_6	Analysis & Vol I	Tao Terence, N. Robert	science	250	Apress	en
b_7	Philosophie der Physik	Heisenberg Werner	science	197	Penguin	de
b_8	Making Software		computer_science	232	O'Reilly	
b_9	Analysis & Vol I	Tao Terence	mathematics	248	HB	en

Figure 6: K_2 sameAs Links (D1)

	title	author	genre	pages	publisher	lang
b_{21}	The Age of Wrath	Eraly Abraham	history	198	Penguin	
b_{22}	Statistical Decision Theory	Pratt John	data_science	236	MIT Press	en, de
b_{23}	The New Machiavelli		fiction	198	Penguin	en
b_{24}	Philosophie der Physik	Heisenberg Werner	science	197	Penguin	de
b_{25}	Analysis & Vol I	Tao Terence	mathematics	248	HB	

Figure 7: K_2 sameAs Links (D2)

2 Part 2: Rule Discovery [8 pts]

2.1 Question 4 [3 pts]

- (a)) What are the two main families of approaches of rule discovery presented in the course and what are the main steps of each of them?
- (b) What is the main challenge that raises for rule discovery under the open world assumption (OWA) and how this challenge is dealt with in AMIE system?

2.1.1 Answer

- (a) Two main families of approaches for rule discovery presented in the course are the top-down and bottom-up strategies. In a top-down approach, one begins with very general rules that cover a large part of the data and then progressively specializes them by adding more atoms to the rule's body. The main steps involve (1) **generating a broad, general rule**, (2) **incrementally specializing the rule by adding conditions**, (3) **evaluating each specialized candidate with quality metrics** (like support and confidence), and (4) **pruning those that do not meet predetermined thresholds**. In contrast, the bottom-up approach starts with specific patterns or candidate rule fragments derived directly from the data and then generalizes or merges them into broader rules. Its steps include (1) **identifying specific, highly reliable rule fragments from observed facts**, (2) **generalizing these fragments by combining them and allowing for slight variations**, and (3) **evaluating and iteratively merging these fragments until high-quality rules emerge**.
- (b) The primary challenge under the open world assumption (OWA) is dealing with incompleteness—namely, the absence of a fact in the knowledge base does not necessarily mean it is false. This lack of negative evidence can lead to an overestimation of counterexamples when evaluating a rule's confidence, thus underestimating the rule's actual quality. The AMIE system addresses this challenge by adopting the Partial Completeness Assumption (PCA). Under PCA, if an entity has at least one recorded fact for a given relation, then it is assumed that the knowledge base is complete with respect to that relation for that entity; hence, missing facts can be treated as genuine negatives. Conversely, if no such fact exists, the entity is not penalized. This PCA-based confidence measure allows AMIE to more accurately reflect the quality of rules in the face of incomplete data.

2.2 Question 5 [3.5 pts]

Let consider the following rules r_1 and r_2 that we assume to be discovered by AMIE tool on the data presented in **Table 8**. The atoms **predicate(x, y)** are written as **(?x predicate ?y)**:

$$r_1 : \textbf{(?b spouse ?a)} \implies \textbf{(?a spouse ?b)}$$
$$r_2 : \textbf{(?a nationality ?b)} \implies \textbf{(?a deathplace ?b)}$$

	spouse	birthplace	deathplace	nationality
#Bob	#Mary	France		France
#Mary		Greece	France	France
#Momo	#Yue	Algeria	Algeria	
#Katsu	#Yori	Senegal	Italy	Italy
#Yue	#Momo	China	China	China
#Yori		Ukraine		France

Figure 8: People Descriptions Dataset

Considering people descriptions presented in **Table 8** compute the support, the standard confidence and the PCA-confidence for r_1 and r_2.

2.2.1 Answer

The first rule is:

$$r_1 : \overbrace{\textbf{(?b spouse ?a)}}^{Body} \implies \overbrace{\textbf{(?a spouse ?b)}}^{Head}$$

From these, the actual triples bodies (?b spouse ?a) in the data are:

(Bob, Mary), **(Momo, Yue)**, **(Katsu, Yori)**, **(Yue, Momo)**

We check whether the **head** "(?a spouse ?b)" also appears in the data:

1. spouse(Mary, Bob): ✗
2. spouse(Yue, Momo): ✓
3. spouse(Yori, Katsu): ✗
4. spouse(Momo, Yue): ✓

So, we have 2 successes out of 4 body matches, then we conclude that the support(r_1) = 2.

The standard confidence follows by the fact that the body is satisfied 4 times, and the head is satisfied 2 times, therefore the standard_conf(r_1) = $\frac{2}{4}$ = 0.5. Finally, we study the PCA-Confidence(r_1). Under the Partial Completeness Assumption, we only treat "missing" head facts as genuine negatives if the subject in the head has at least one known spouse triple. Concretely:

1. spouse(Bob, Mary) ✓ and as spouse(Mary, Bob) is not present, we dont count it as negative.
2. spouse(Momo, Yue) ✓ and spouse(Yue, Momo) ✓, therefore this is a sucess rule inference.
3. spouse(Katsu, Yori) ✓ and as spouse(Yori, Katsy) is not present, we don't count it as negative.

Then, there are no counter examples under PCA, which means that the confidence is:

$$\text{PCA-Confidence}(r_1) = \frac{\text{support}(r_1)}{\text{support}(r_1) + cex_{PCA}} = \frac{2}{2+0} = 1$$

The second rule is:

$$r_2 : \overbrace{\textbf{(?a nationality ?b)}}^{Body} \implies \overbrace{\textbf{(?a deathplace ?b)}}^{Head}$$

From these, the actual triples bodies (?a nationality ?b) in the data are:

(Bob, France), **(Mary, France)**, **(Katsu, Italy)**, **(Yue, China)**, **(Yori, France)**

We check whether the **head** "(?a deathplace ?b)" also appears in the data:

1. deathplace(Bob, France): ✗
2. deathplace(Mary, France): ✓
3. deathplace(Katsu, Italy): ✓
4. deathplace(Yue, China): ✓
5. deathplace(Yori, France): ✗

So, we have 3 successes out of 5 body matches, then we conclude that the support(r_2) = 3.

The standard confidence follows by the fact that the body is satisfied 5 times, and the head is satisfied 3 times, therefore the standard_conf(r_5) = $\frac{3}{5}$ = 0.6. Finally, we study the PCA-Confidence(r_2). Under the Partial Completeness Assumption, we only treat "missing" head facts as genuine negatives if the subject in the head has at least one known deathplace triple. Concretely:

1. nationality(Bob, France) ✓ and as deathplace(Bob, France) is not present, we dont count it as negative.
2. nationality(Mary, France) ✓ and deathplace(Mary, France) ✓, therefore this is a sucess rule inference.
3. nationality(Katsu, Italy) ✓ and as deathplace(Katsu, Katsy) ✓, therefore this is another sucess rule inference.
4. nationality(Yue, China) ✓ and as deathplace(Yue, China) ✓, therefore this is a sucess rule inference.
5. nationality(Yori, France) ✓ and as deathplace(Yori, France) is not present, we dont coun't it as negative.

Then, there are no counter examples under PCA, which means that the confidence is:

$$\text{PCA-Confidence}(r_2) = \frac{\text{support}(r_2)}{\text{support}(r_2)+cex_{PCA}} = \frac{3}{3+0} = 1$$

2.3 Question 6 [1.5 pts]

Let us consider the following rule:

$$r_3\text{: (?a birthplace ?b) and (?a deathplace ?b)} \implies \text{(?a nationality ?b)}$$

that is discovered on data described in **Table 8**. Give the SPARQL query that translates the rule r_3 for allowing using it for predicting new facts for **nationality** predicate.

2.3.1 Answer

A convenient way to express the rule in SPARQL is via a CONSTRUCT query that finds all **subjects** ?a and **objects** ?b satisfying both birthplace and deathplace, while ensuring ?a nationality ?b does not already exist. For instance:

```
CONSTRUCT {
?a :nationality ?b
}
WHERE {
  ?a :birthplace ?b .
  ?a :deathplace ?b .
  FILTER NOT EXISTS {
    ?a :nationality ?b
  }
}
```

This query "discovers" new triples (?a :nationality ?b) whenever both (?a :birthplace ?b) and (?a :deathplace ?b) are present in the data.