Knowledge Discovery in Graph data – exam[*]

Printed documents and notes are allowed –duration: 2:00

Computer Science Master 2 – Data Science – Paris Saclay University

February 14, 2025

In what follows, the data are presented in tables and not in graphs for readability reasons. The triple subjects are in the first column, the predicates are in the first line and the objects are in the other cells of each table.

1 Key discovery for data linking *[12pts]*

Question 1. [2.5pts] 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.

Question 2. [2.5pts] 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?

Question 3. 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]*

[*] The mark scale is given as an indication.

	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

Table 1: Extract of book descriptions (D1)

(d) Let consider a key $K_1 = \mathit{hasKey}(\mathit{Book})(\mathit{title}, \mathit{hasAuthor})$ and $K_2 = \mathit{hasKey}(\mathit{Book})(\mathit{hasAuthor}, \mathit{lang})$.

Let D_2 be a second dataset given in Table 2. What would be the `sameAs` links that can be inferred when applying K_1 and K_2 to $D_1 \times D_2$ 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	

Table 2: Book descriptions (D2)

2 Part 2: Rule Discovery

Question 4. [3pts]

(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?

Question 5.[3.5pts] 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 3. The atoms $\mathit{predicate}(x, y)$ are written as `(?x predicate ?y)`:

- `r1: (?b spouse ?a) => (?a spouse ?b)`
- `r2: (?a nationality ?b) => (?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

Table 3: People descriptions

Considering people descriptions presented in Table 3 compute the support, the standard confidence and the PCA-confidence for r_1 and r_2.

Question 6.[1.5pts]

Let us consider

```
r3:  (?a birthplace ?b) and (?a deathplace ?b) => (?a nationality ?b)
```

be a rule that is discovered on data described in Table 3.

Give the SPARQL query that translates the rule r_3 for allowing using it for predicting new facts for `nationality` predicate.