Tutorial on Semantic Schema Discovery: principles, methods and future research directions Part 3

Kenza Kellou-Menouer, Nikolaos Kardoulakis, Georgia Troullinou, Zoubida Kedad, Dimitris Plexousakis, Haridimos Kondylakis

Équipes Traitement de l'Information et Systèmes

ROADMAP

Introduction
& Applications
(30 min)

1

Kenza
Kellou-Menouer

Pattern Discovery
(25 min)

3

Open Issues
(15 min)

5

2

Dimensions of Analysis
& Implicit Schema
Discovery
(35 min)

4

Explicit Schema
Enrichment
(20 min)

6

Hands-on
(20 mins)

Pattern Discovery Approaches

Terminology

Pattern:

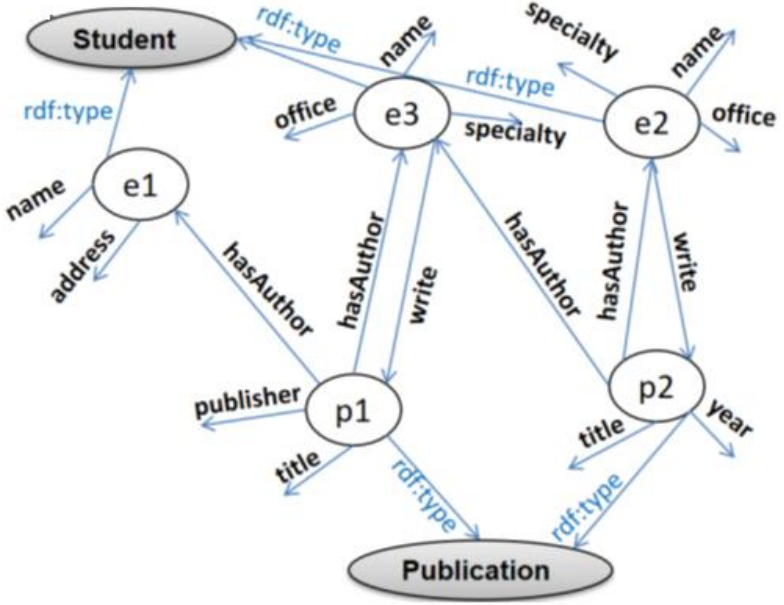
A set of properties V describing instances of a type.

e.g: $V = \{\overrightarrow{name}, \overrightarrow{address}, \overleftarrow{hasAuthor}\}$

Pattern Discovery Approaches:

Analyzing the co-occurrence relationships among the properties in the dataset.

- Identifying the possible set of properties that could describe an instance of a type.



Terminology

Pattern:

A set of properties V describing instances of a type:

- $V = \{p_1, p_2, ..., p_n\}$

Exact Pattern:

V contains all the properties p describing an instance e

– ∀ p ∈ V : p describes e;

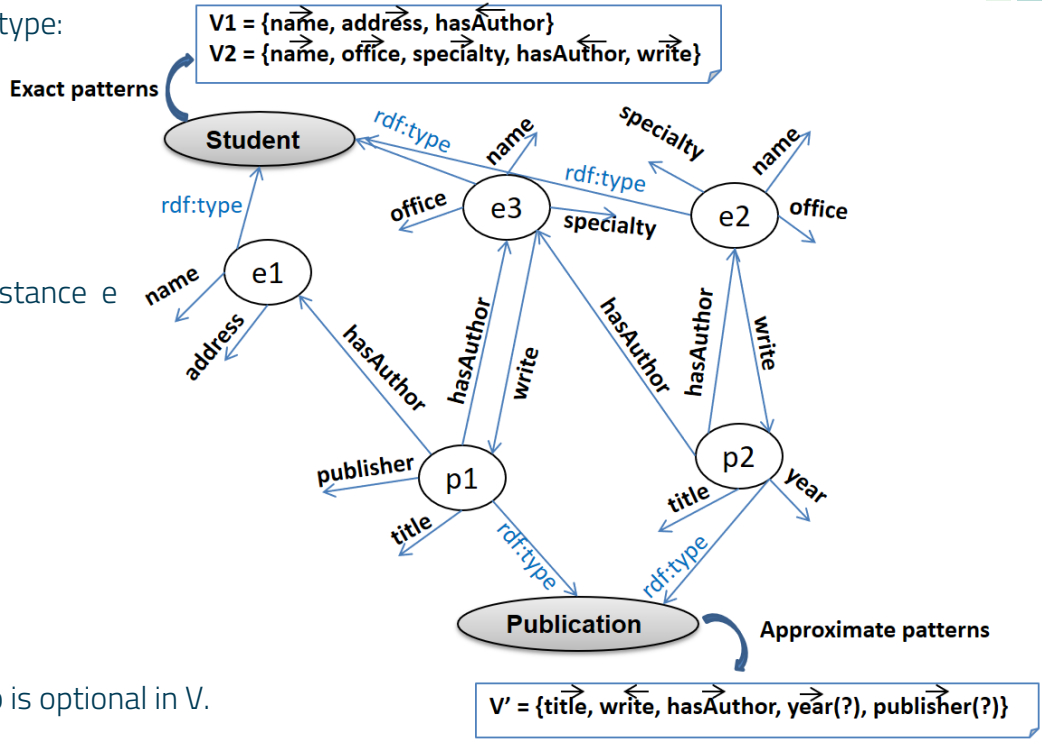
– ∀ p describing e: p ∈ V .

Approximate Pattern:

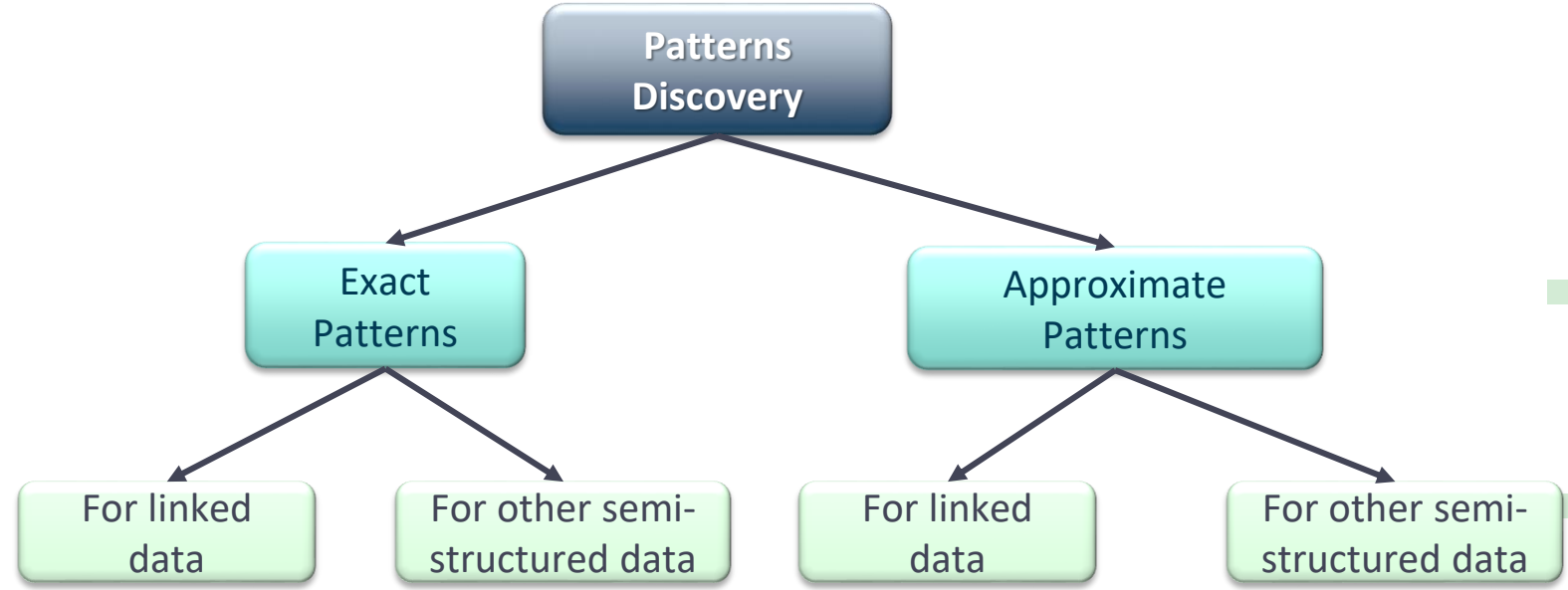
Describe a set of similar instances T

– ∀ p describing all the instances in T: p ∈ V ;

– If p describe only some instances in T, then p is optional in V.



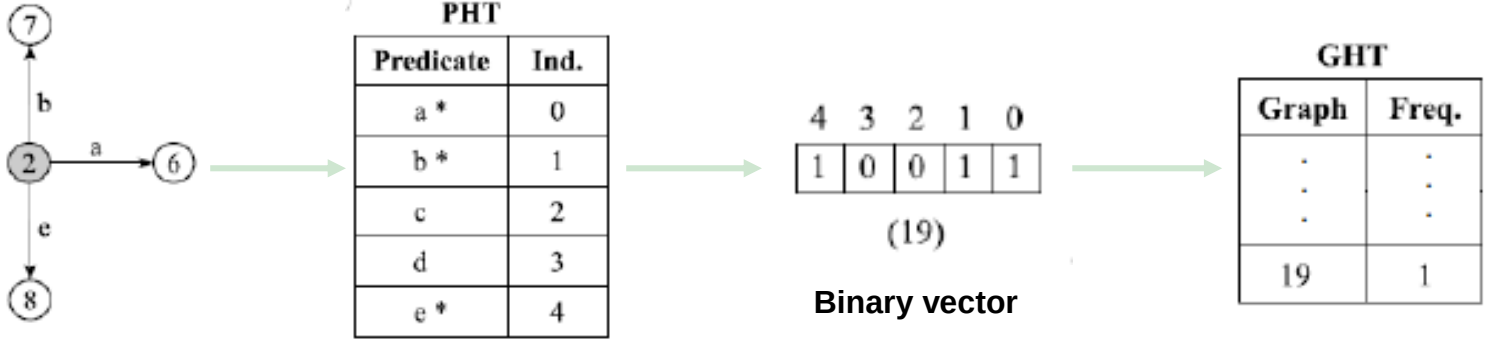
Pattern discovery approaches classification



Exact Pattern Discovery Approaches

For linked data:

Approach	Inputs		Techniques	Outputs	Pros & Cons
	Data Description	Setting			
Belghaouti et al. *FreGraPad* (2016) [1]	RDF data: stream, outgoing properties	No required setting	Statistics by browsing and counting instances	Exact patterns with the frequency of each one	+ Incremental



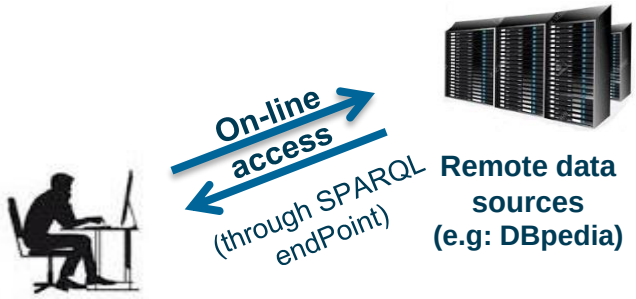
Graph n: set of triples

[1] Belghaouti, F., Bouzeghoub, A., Kazi-Aoul, Z., Chiky, R. Fregrapad: Frequent rdf graph patterns detection for semantic data streams. In: Research Challenges in Information Science (RCIS), 2016

Exact Pattern Discovery Approaches

For linked data:

Approach	Inputs		Techniques	Outputs	Pros & Cons
	Data Description	Setting			
Kellou-Menouer et Kedad *SchemaDecrypt(++)* (2017 - 2020) [2, 2']	**RDF(S) data:** remote source with restricted access, incoming and outgoing properties, part of schema (*rdf:type*)	No required setting	Statistics on the properties of the instances. The set of properties could be automatically extracted relying on RDFS statements.	Exact patterns with the frequency of each one	+ On-line approach + Parallelized + Scalability can be by individual source restrictions



Source restrictions:

- limited size of answer
- limited time to answer a query
- limited number of queries

[2] Kellou-Menouer, K., Kedad, Z.: On-line versioned schema inference for large semantic web data sources. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management, SSDBM 2017.
[2'] Kellou-Menouer, K., Kedad, Z.: SchemaDecrypt++: Parallel on-line versioned schema inference for large semantic web data sources. Information Systems Journal 2020.

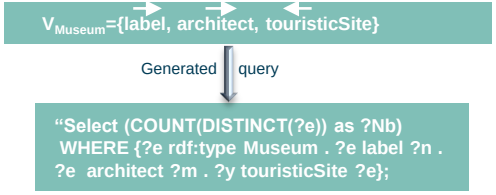
Exact Pattern Discovery Approaches

For linked data:

Approach	Inputs		Techniques	Outputs	Pros & Cons
	Data Description	Setting			
Kellou-Menouer et Kedad *SchemaDecrypt(++)* (2017 - 2020) [2, 2']	**RDF(S) data:** remote source with restricted access, incoming and outgoing properties, part of schema (*rdf:type*)	No required setting	Statistics on the properties of the instances. The set of properties could be automatically extracted relying on RDFS statements.	Exact patterns with the frequency of each one	+ On-line approach + Parallelized + Scalability can be by individual source restrictions

Basic Idea:

- The candidate versions for a class: all the combinations of its properties
- Each combination / candidate version is tested by sending the corresponding query to the source



A combinatorial problem:

- 2^n queries to execute for a class with n properties
- Average number of properties in DBpedia is 150:
  - 2^{150} queries to send
  - 15 queries by second(1)

  => 10^{36} years(2)

- If the number of instances is positive, the candidate version is validated

(1) The DBpedia online server can test a maximum of 15 queries per second.

(2) $2^{150}/15$ seconds ≈ 2^{146} seconds, knowing that 1 year ≈ 31 536 000 seconds ≈ 2^{25} seconds, 2^{146} seconds ≈ 2^{121} years ≈ 10^{36} years.

Exact Pattern Discovery Approaches

For linked data:

Approach	Inputs		Techniques	Outputs	Pros & Cons
	Data Description	Setting			
Kellou-Menouer et Kedad *SchemaDecrypt(++)* (2017 - 2020) [2, 2']	**RDF(S) data:** remote source with restricted access, incoming and outgoing properties, part of schema (*rdf:type*)	No required setting	Statistics on the properties of the instances. The set of properties could be automatically extracted relying on RDFS statements.	Exact patterns with the frequency of each one	+ On-line approach + Parallelized + Scalability can be by individual source restrictions

Strategy:

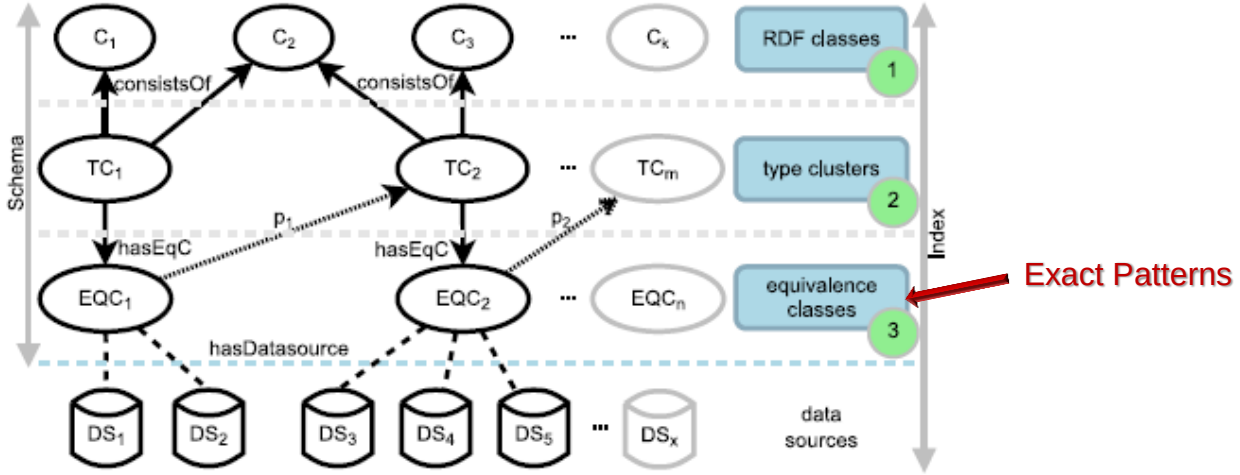
- Probabilistic profile: A property vector containing the properties of the instances of the class and their probabilities
  - Testing the most probable patterns at first
  - Reducing the search space
  - Defining a stopping criteria
- Properties co-occurrence rules
  - Pruning the search space
- Parallelization of candidate patterns exploration (SchemaDecrypt++)

[2] Kellou-Menouer, K., Kedad, Z.: On-line versioned schema inference for large semantic web data sources. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management, SSDBM 2017.
[2'] Kellou-Menouer, K., Kedad, Z.: SchemaDecrypt++: Parallel on-line versioned schema inference for large semantic web data sources. Information Systems Journal 2020.

Exact Pattern Discovery Approaches

For linked data:

Approach	Inputs		Techniques	Outputs	Pros & Cons
	Data Description	Setting			
Konrath et al. *SchemEx* (2012) [3]	**RDF data:** distributed sources with unrestricted access, outgoing properties, part of schema (*rdf:type*)	No required setting	Formal Method by browsing the data to build three layers	Exact patterns and an index with Three layers	+ Can be applied to RDF stream data + Web scale



[3] Konrath, M., Gottron, T., Staab, S., Scherp, A.: Schemex: efficient construction of a data catalogue by stream-based indexing of linked data. Web Semantics: Science, Services and Agents on the World Wide Web (2012)

Exact Pattern Discovery Approaches

For other semi-structured data:

Approach	Inputs		Techniques	Outputs	Pros & Cons
	Data Description	Setting			
Ruiz et al. (2015) [4]	**JSON data:** outgoing properties, the kind of the object	No required setting	Formal Method by dividing the objects on the basis of the kind (type) with *Map-Reduce*	Exact patterns expressed in JSON	+ Provide tools for schema viewing and validation + Scalable

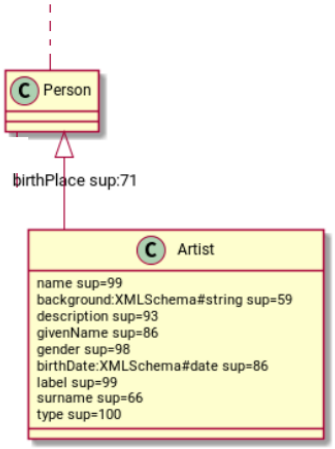
```
Entity Journal {
  Version 1 {
    issn: Tuple [String, String]
    name: String
    discipline: String
  }
  Version 2 {
    issn: Tuple [String, String]
    name: String
    discipline: String
    number: int
  }
}
```


[4] Ruiz, D.S., Morales, S.F., Molina, J.G.: Inferring versioned schemas from NoSQL databases and its applications. In: International Conference on Conceptual Modeling, pp. 467–480. Springer (2015)

Approximate Pattern Discovery Approaches

For linked data:

Approach	Inputs		Techniques	Outputs	Pros & Cons
	Data Description	Setting			
Issa S. et al. *LOD-CM* (2019) [5]	**RDF data:** outgoing properties, part of schema (*rdf:type*)	Threshold for classifying a pattern as frequent	Statistics and Machine learning (FP-Growth) by mining maximal frequent patterns upon properties	Approximate patterns	+ Takes into account user quality expectations − Iterative process

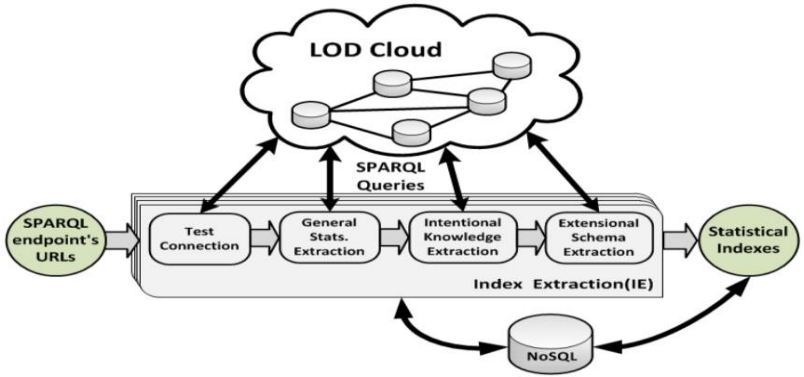


[5] Issa, S., Paris, P., Hamdi, F., Cherfi, S.S.: Revealing the conceptual schemas of RDF datasets. In: Advanced Information Systems Engineering - 31st International Conference, CAiSE 2019.

Approximate Pattern Discovery Approaches

For linked data:

Approach	Inputs		Techniques	Outputs	Pros & Cons
	Data Description	Setting			
Benedetti et al. *LODeX* (2014-2015) [6, 6']	RDF data: *rdf:type* exploited if provided	No required setting	Statistics	Approximate patterns expressed on an RDF graph	+ Visual representation − Relies on online SPARQL endpoints with scalability restrictions



[6] Benedetti, F., Bergamaschi, S., Po, L.: Online index extraction from linked open data sources. Proceedings of the Second International Workshop on Linked Data for Information Extraction (LD4IE 2014) co-located with the 13th International Semantic Web Conference (ISWC 2014)
[6'] Benedetti, F., Bergamaschi, S., Po, L.: Exposing the underlying schema of LOD sources. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015, Singapore, December

Approximate Pattern Discovery Approaches

For linked data:

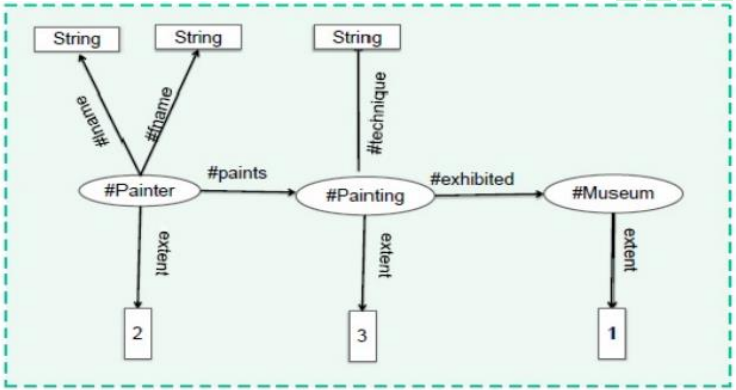
Approach	Inputs		Techniques	Outputs	Pros & Cons
	Data Description	Setting			
Zneika et al. (2016) [7]	**RDF data:** In/outgoing properties, part of schema (*rdf:type* exploited if provided)	Settings for cost function	Machine learning using a pattern mining algorithm (PaNDa+)	Approximate patterns expressed on an RDF graph with the occurrence number of each pattern (using "extent" statement)	+ Detects the most representative approximate patterns − Requires dataset transformation into a binary matrix

	Painter(c)	Painting(c)	lname	fname	paints	exhibited	R_paints	R_exhibited	Museum(c)
Picasso	1	0	1	1	1	0	0	0	0
Rembrant	1	0	1	1	1	0	0	0	0
Woman	0	1	0	0	0	0	1	0	0
Guernica	0	1	0	0	0	1	1	0	0
Abraham	0	1	0	0	0	1	1	0	0
museum.es	0	0	0	0	0	0	0	1	1

Binary matrix

ID	Pattern	Correspondence class
P1	Painting(c),exhibited, revers_paint	3
P2	Painter(c),paints, fname, lname	2
P3	Museum(c)	1

Extracted patterns



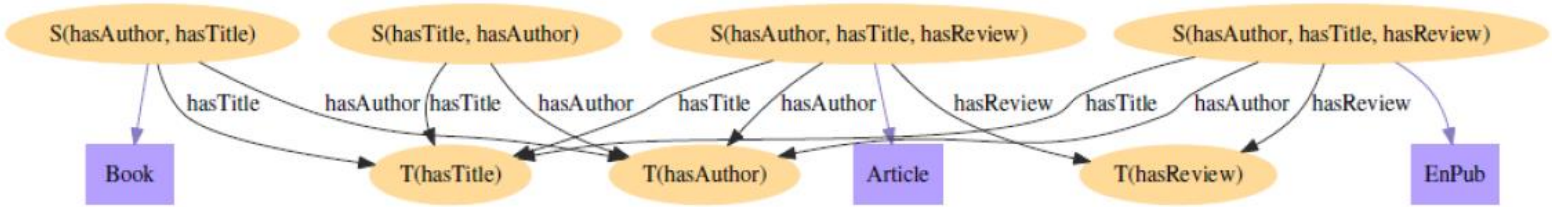
RDF Summary graph for the set of patterns

[7] Zneika, M., Lucchese, C., Vodislav, D., Kotzinos, D.: Summarizing linked data RDF graphs using approximate graph pattern mining. In: Proceedings of the 19th International Conference on Extending Database Technology, EDBT 2016.

Approximate Pattern Discovery Approaches

For linked data:

Approach	Inputs		Techniques	Outputs	Pros & Cons
	Data Description	Setting			
Cebiric et al. (2015) [8]	**RDFS data:** outgoing properties, part of schema (*rdf:type*, *rdfs:domain*, *rdfs:range*, *rdfs:subClassOf* exploited if provided)	No required setting	Formal Methods by grouping properties according to the type of their subject in the triples, by considering *rdf:type* statements	Approximate patterns expressed on an RDF Query-oriented graph	+ Considers both implicit and explicit triples − Computationally expensive $O(\|G^5\|)$ − The optional properties are not marked in the produces schema



[8] Cebiric, Š., Goasdoué, F., Manolescu, I.: Query-oriented summarization of rdf graphs. Proceedings of the VLDB Endowment 8(12), (2015)

Approximate Pattern Discovery Approaches

For other semi-structured data:

Approach	Inputs		Techniques	Outputs	Pros & Cons
	Data Description	Setting			
Baazizi et al. (2017-2019) [9, 9']	**JSON data:** outgoing properties, the kind of the object	Precision and conciseness level could be setting	Formal Method by a random division of data with *Map-Reduce*	Approximate patterns described by regular expressions	+ Spark-based scalable approach + Incremental

Dataset:

1.Person { id : 12, name : John Smith, age : 14}
2.Office { id : 31, number : 3, address : 4 rue armengaud}
3.Person { id : 4, name : Kenza Kellou-Menouer}

1. Map: Structure inference

T_1 = { id : Number, name : String, age : Number}
T_2 = { id : Number, number : Number, address : String}
T_3 = { id : Number, name : String}

2. Reduce: Type merging (Approximate Patterns)

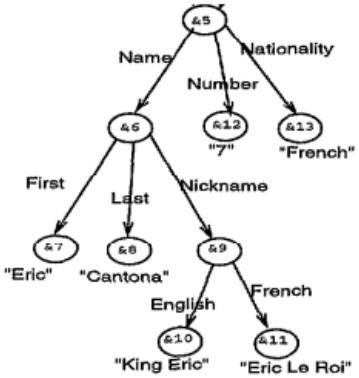
$T_{1,3}$ = { id : Number, name : String, (age : Number) ?}
T_2 = { id : Number, number : Number, address : String}

[9] Baazizi, M.A., Colazzo, D., Ghelli, G., Sartiani, C.: Parametric schema inference for massive JSON datasets. VLDB J. 28(4), 497–521 (2019)
[9'] Baazizi, M.A., Lahmar, H.B., Colazzo, D., Ghelli, G., Sartiani, C.: Schema inference for massive JSON datasets. In: Proceedings of the 20th International Conference on Extending Database Technology, EDBT 2017.

Approximate Pattern Discovery Approaches

For other semi-structured data:

Approach	Inputs		Techniques	Outputs	Pros & Cons
	Data Description	Setting			
K. Wang et H. Liu. (1997) [10]	OEM data: outgoing properties	Support threshold for pattern mining	Machine learning using a pattern mining algorithm	Approximate patterns schema	+ Detects the most representatifs patterns − Optional properties are not included in the approximate pattern



[10] Wang, K., Liu, H.: Schema discovery for semistructured data. In: Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD- 1997)

Rewiew of Pattern Discovery Approaches

- A set of approaches proposed for:
  - Exact patterns discovery [1-4]
  - Approximate patterns discovery [5-10]

- In an approximate pattern, the co-occurrence of the properties is not clearly indicated:
  - Specifying the percentage of each property among the instances of a class [5]
  - Marking of the optional properties of a pattern [9, 9'], or not even [8, 10]
  - Including only the no optional properties in the pattern [7].

- Different data sources are processed:
  - Streaming data [1]
  - Remote data with constrained access [2, 2', 6, 6']
  - Distributed data [3]
  - Local data [4-7, 8-10]

Break.
To be continued...