Foundations of Data and Knowledge Bases

Structured Data Collections:
Tables, Graphs, Unranked Trees, Data Trees, Relational Structures

Joachim Niehren

Links: Linking Dynamic Data
Inria Lille

November 10, 2020

Outline

Database tables

Relational schema

alphabet Δ for words in tables is finite set
ranked alphabet Γ of relation symbols

Database = Relations between words

for all relation symbols γ or arity k a k-ary relation:

$$R_\gamma \subseteq \underbrace{\Delta^* \times \ldots \times \Delta^*}_{k\ \ times}$$

Example

table of (first-names, last names) pairs of PhD students is relation

$$R_{first-name\_last-name} = \{(``tom'','' sebastian''), (``antoine'','' ndione'')\}$$

Outline

Colored Digraphs

Directed Graphs (Digraphs)

nodes set: V
edge set: $E \subseteq V \times V$
digraph: $G = (V, E)$

Color Alphabets

C_{nod} set of node colors
C_{edg} set of edge colors

Colored Digraph

digraph (V, E)
for every node color $c \in C_{nod}$ a set $V_c \subseteq V$
for every edge color $c \in C_{edg}$ a set $E_c \subseteq E$
colored digraph $G = (V, E, (V_c)_{c \in C_{nod}}, (E_c)_{c \in C_{edg}})$

Reachability

Reachable Nodes

v_2 is reachable from v_1 over a path in G iff $(v_1, v_2) \in E^*$
How to define reachability over paths of edges with color a?
Can you define what a path is?

Example

$V = \{Li, Pa, Ly, Ma\}$
$E = \{(Li, Pa), (Pa, Ly), (Ly, Ma)\}$
Ma is reachable from Li in this colored digraph, but Li is not reachable from Ma there.

Terms as Digraphs

Colors

Colors of nodes in Σ
Colors of edges in $\{i \mid f \in \Sigma, 1 \leq i \leq ar(f)\}$
$graph(t) = (nod(t), edg(t), (lab_a(t))_{a \in \Sigma}, (pos_i(t))_{i \in \mathbb{N}})$

Nodes and Edges $nod(t) \subseteq \mathbb{N}^*$

$$nod(a) = \{\epsilon\}$$
$$nod(f(t_1, \ldots, t_n)) = \{\epsilon\} \cup \cup_{i=1}^{n} i \cdot nod(t_i)$$
$$edg(a) = \emptyset$$
$$edg(f(t_1, \ldots, t_n)) = \{(\epsilon, i) \mid 1 \leq i \leq n\}$$
$$\cup \cup_{i=1}^{n} \{(i{\cdot}\pi, i{\cdot}\pi') \mid (\pi, \pi') \in edg(t_i)\}$$

Exercise: Node Colors

1. For every node color $a \in \Sigma$ define sets $lab_a(t)$ of nodes of t whose label is a.

Relational Databases

Tables

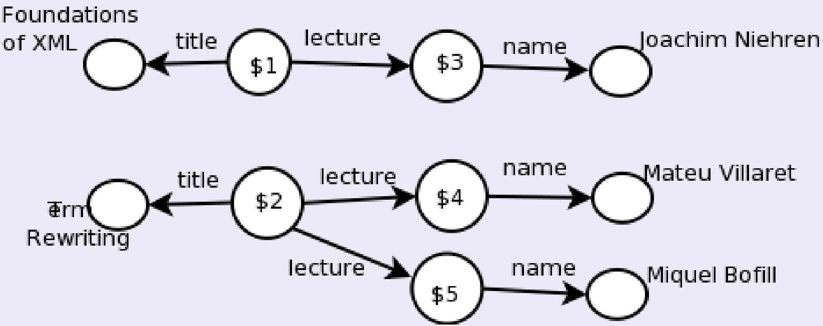
lecture	title
\$1	Foundations of XML
\$2	Term Rewriting

teacher	lecture
\$3	\$1
\$4	\$2
\$5	\$2

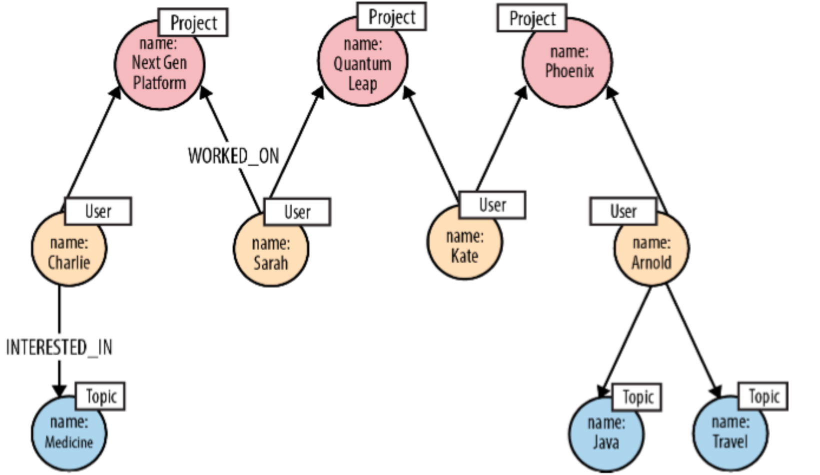
teacher	name
\$3	Joachim Niehren
\$4	Mateu Villaret
\$5	Miquel Bofill

Graphical Representation

Colored Digraph



Exercise: Graph Databases as Colored Digraphs



2. How can you formalize the above graph database as a colored digraph?

Exercise: Paths in Graph Databases

3. Let us call two users u and u' related in the above database if there is a sequence of users $u_0, \ldots, u_n$ with $u = u_0$ and $u' = u_n$ such that for all $1 \leq i \leq n$ the users u_i and u_{i-1} worked on the same topic. Can you write a regular expression based on edge colors and inverse edge colors, so that all related users are linked by a path recognized by your regular expression?

Outline

Signature=Vocabulary

Unranked Signature Σ

- a set of symbols
- no arities

Examples for unranked tree over $\{a, b\}$

$a(b, b, b, b, b, b)$
$a(b(a(b(a(a, b)))))$
$a(b(a, b), a, b(b, a(b)))$

Inductive Definition of Unranked Trees

Parameter

Σ an unranked signature

Set of unranked trees $T_\Sigma^{\leq m}$ of depth $\leq m$

$$
\begin{aligned}
T_\Sigma^{\leq 0} &= \emptyset \\
T_\Sigma^{\leq m+1} &= \{a(t_1, \ldots, t_n) \mid a \in \Sigma, t_1, \ldots, t_n \in T_\Sigma^{\leq m},\ n \geq 0\} \cup T_\Sigma^{\leq m}
\end{aligned}
$$

Set of all unranked trees

$$T_\Sigma = \cup_{m=0}^{\infty} T_\Sigma^m$$

Equivalent Definitions

Recursive

T_Σ is the least set that contains all pairs $a(t_1, \ldots, t_n)$ where $a \in \Sigma$ and $(t_1, \ldots, t_n) \in (T_\Sigma)^n$ for some $n \geq 0$:

Mathematical

$$T_\Sigma \quad = \quad \uplus_{n \geq 0} \Sigma \times (T_\Sigma)^n$$

Backus-Naur form (BNF)

$t \in T_\Sigma ::= a(t_1, \ldots, t_n)$ where $n \geq 0$.

Example

signature $\Sigma = \{bib, author, book, title\}$
unranked trees in T_Σ: $bib(book(author, author, title), book(author, title))$
how can we define the schema of admissible bibliographies?

Unranked Trees as Digraphs

Nodes and Edges

$$nod(t) \subseteq \mathbb{N}^*$$
$$nod(a(t_1, \ldots, t_n)) = \{\epsilon\} \cup \cup_{i=1}^{n} i \cdot nod(t_i)$$

father-child edges in analogy to terms

Colors

Node colors analoguous as for terms: Σ
How many edge colors does one need? Are finitely many edge colors sufficient, if Σ is finite?

Words as Unranked Trees

Flat trees

fix symbol for the root $r \in \Sigma$

transformation from Σ^* to T_Σ: $a{\cdot}b{\cdot}c \Rightarrow r(a, b, c)$

Deep trees

fix symbol for the end of a word nil

transformation from Σ^* to $T_{\Sigma \cup \{nil\}}$: $a{\cdot}b{\cdot}c \Rightarrow a(b(c(nil)))$

Linearization of Unranked Trees as Words

Preorder traversal as in XML

$bib(book(author("Abiteboul"), author("Hull") \ldots title("FoXML"))) \Rightarrow$
$< bib >< book >< author > Abiteboul < /author >< author > Hull <$
$author/ > \ldots < title > FoXML < /title >< /book >< /bib >$

Exercise

4. Find a deterministic finite automaton for the regular expression $(a^* + b^*) \cdot (c^* + d^*)$.

Outline

Relational Structures

Relational signature $\Sigma = \mathrm{Consts} \uplus \mathrm{Rels}$

- Consts is a set of constants (that may be infinite, for instance the set of data values i.e. the set of strings of a finite alphabet.
- Rels is a finite set of relation symbols. Every symbols has a fixed arity given by a function $\mathrm{ar} : \mathrm{Rels} \to \mathbb{N}$.

Finite Σ-Structures $S = (\mathrm{Dom}, .^S)$

- Dom is a finite set of elements called the domain (it may contain the nodes of a graph and a finite subset of data values.)
- $.^S$ gives interpreation to all symbols of Σ:

$$\begin{array}{ll} a^S \in \mathrm{Dom} & \text{for } a \in \mathrm{Consts} \\ r^S \subseteq \mathrm{Dom}^{\mathrm{ar}(r)} & \text{for } r \in \mathrm{Rels} \end{array}$$

Example Database

lecture	location	title	teacher
$1	Girona	Foundations of XML	$3
$2	Girona	Term Rewriting	$4
$2	Girona	Term Rewriting	$5

teacher	fname	lname
$3	Joachim	Niehren
$4	Mateu	Villaret
$5	Miquel	Bofill

Relational Structure

$$
\begin{aligned}
&\mathrm{Rels} = \{\mathit{lecture}, \mathit{teacher}\}\\
&\mathrm{ar}(\mathit{lecture}) = 4\\
&\mathrm{ar}(\mathit{teacher}) = 3\\
&\mathrm{Consts} = \{\text{“Girona”}, \text{”Joachim”}, \text{”Niehren”}, \text{”Mateu”}, \text{”Villaret”},\\
&\qquad\qquad \text{”Miquel”},\text{”Bofill”}, \text{“Foundations of XML”},\\
&\qquad\qquad \text{“Term Rewriting”}\}\\
&\mathrm{Dom} = \{\$1, \$2, \$3, \$4, \$5\} \uplus \mathrm{Consts}\\
&\mathit{lecture}^S = \{(\$1, \text{”Girona”},\text{”Foundations of XML”}, \$3),\\
&\qquad\qquad (\$2, \text{”Girona”},\text{”Term Rewriting”}, \$4),\\
&\qquad\qquad (\$2, \text{”Girona”},\text{”Term Rewriting”}, \$5)\}\\
&\mathit{teacher}^S = \{(\$3, \text{”Joachim”},\text{”Niehren”}),\\
&\qquad\qquad (\$4, \text{”Mateu”},\text{”Villaret”}),\\
&\qquad\qquad (\$5, \text{”Miquel”},\text{”Bofill”})\}\\
&a^S = a \quad \text{for all } a \in \mathrm{Consts}
\end{aligned}
$$

Relational Structure for Words

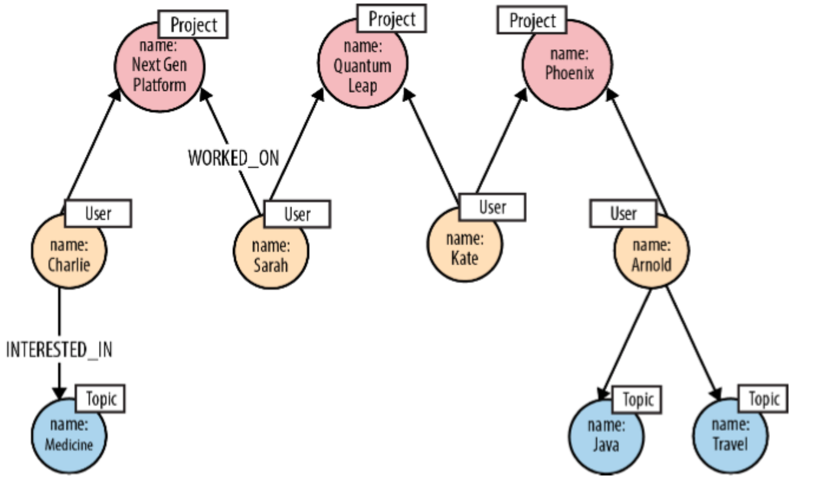
Signature

alphabet :	Δ
constants:	$\mathrm{Consts} = \{\mathrm{start}, \mathrm{last}\}$
relation symbols:	$\mathrm{Rels} = \{\mathrm{next}^*\} \cup \{lab_a \mid a \in \Delta\}$

Relational structure for word $w = a_1 \dots a_n \in \Delta^*$

domain	$\mathrm{Dom}^w = \{0, \dots, n\}$ (is always non-empty)	
constants	$\mathrm{start}^w = 0$	
	$\mathrm{last}^w = n$	
predicates	$lab_a^w = \{i \mid a_i = a\}$	where $a \in \Delta$
	$(\mathrm{next}^*)^w = \{(i,j) \mid 0 \leq i \leq j \leq n\}$	

Exercise



5. Can you formalize the above graph database as a relational structure?

Outline

Data Trees

Data values are words in Γ^* for some finite set Γ

Unranked Data Trees

Data trees are unranked trees t with mapping $text : nodes(t) \rightarrow \Gamma^*$.

What can one do with data values?

Test for equality
Select data value
Apply string operations
Convert to numbers
Arithmetics

Data Graphs

Exercise

What could be a data word?
What could be a colored data graph?

Question

Are relationional structures with additional data stilll relational structures?