


Semi-structured data

Patrick Lambrix
Department of Computer and Information Science
Linköpings universitet


1



Semi-structured data

- Data is not just text, but is not as well-structured as data in databases
- Occurs often in web databanks
- Occurs often in integration of databanks


2



Semi-structured data - properties

- irregular structure
- implicit structure
- partial structure
- a posteriori 'data guide' versus a priori schema
- large data guides


3



Semi-structured data - properties

- It should be possible to ignore the data guide upon querying
- Data guide changes fast
- object can change type/class
- difference between data guide and data is blurred


4



Semi-structured data - model

- network of nodes
- object model (oid)
- query: path search in the network

5

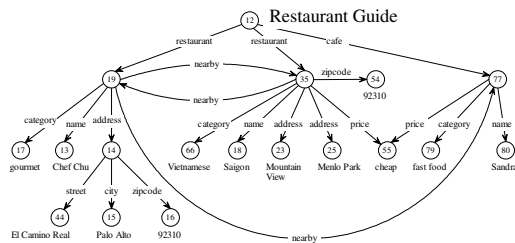


OEM (Object Exchange Model)

- Graph
- Nodes: objects
 - oid
 - atomic or complex
 - atoms: integer, string, gif, html, ...
 - value of a complex object is a set of object references (label, oid)
- Edges have labels
- OEM is used by a number of systems (ex. Lorel)

6

OEM example



7

Lorel query language

- Find all places to eat Vietnamese food

```
select P
from RestaurantGuide.% P
where P.category grep "ietnamese"
```
- Find the names and streets of all restaurants in Palo Alto

```
select R.name, A.street
from RestaurantGuide.restaurant{R}.address A
where A.city = "Palo Alto"
```

8

Lorel query language

- Find all restaurants to eat with zipcode 92310

```
select RestaurantGuide.restaurant
where
RestaurantGuide.restaurant(.address)?.zipcode = 92310
```

Wildcards and variables

- | | |
|---------------------|----------------------------|
| ? - 0 or 1 path | - object variables |
| + - 1 or more paths | select P from Guide.% P |
| * - 0 or more paths | select A from #.address{A} |
| # - any path | - path variables |
| % - 0 or more chars | select Guide.# @P.name |

9

Data Guides

- A structural summary over a data source that is used as a dynamic schema
- Is used in query formulation and optimization
- Is often created a posteriori
- Properties:
 - concise
 - accurate
 - convenient

10

Data Guides - definitions

- Label path: sequence of labels
 $L1.L2. \dots .Ln$
- Data path: alternating sequence of labels and oid:s
 $L1.o1.L2.o2. \dots .Ln.on$
- Data path d is an instance of label path l if the sequences of labels are identical in l and d .

11

Data Guides - definitions

- A data guide for object s is an object d such that every label path of s has exact one data path instance in d , and each label path in d is a label path of s .

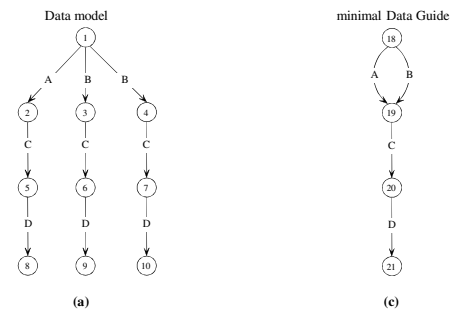
12

Data Guides

- A data source can have several data guides
- Minimal data guides
the smallest data guides

13

Data Guides - example



14

Minimal Data Guides

- Concise
- May be hard to maintain
Example: child node for 10 with label E

15

Strong Data Guides

Intuitively:

"label paths that reach the same set of objects in the data model = label paths that reach the same objects in the data guide"

16

Strong Data Guides - definitions

An object o can be reached from s via l if there is a data path of s that is an instance of l and that has o as last oid ($L1.o1.L2.o2. \dots Ln.o$)

The target set for label path l in object s is the set of objects that can be reached from s via l . Notation: $T(s,l)$

$L(s,l)$: set of label paths of s that have the same target set in s as l .

17

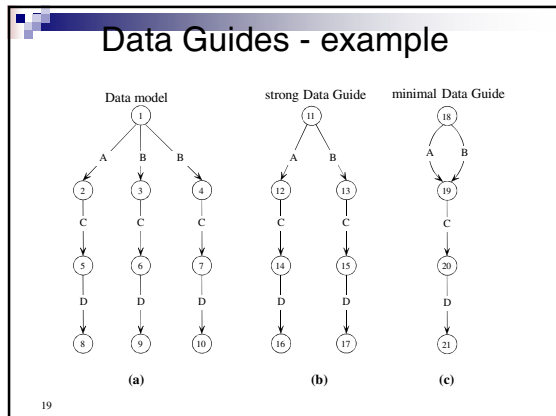
Strong Data Guides - definitions

Definition:

d is a strong data guide for s if for all label paths l of s it holds that $L(s,l) = L(d,l)$

There is a 1-1-mapping between target sets in the data model and nodes in a strong data guide.

18



Strong Data Guides - algorithm

Implementation:

- Traverse data model depth-first.
- Each time you find a new target set for label path l , create a new object in the data guide.

If the target set is already represented in the data guide, do not create a new object, but link to the existing object.

20

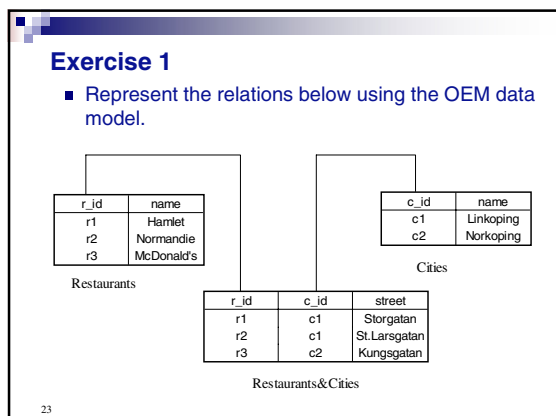
Strong Data Guides - use

- Easier to maintain
- Used as path index for query optimization

21

Semi-structured data - exercises

22



Exercise 2

■ Using the data model from the previous question, formulate the following queries using Lorel:

- find all the restaurants that are located in Linköping
- find the address (city and street) of the "Hamlet" restaurant
- list the restaurants by city (equivalent of GROUP BY)

24

