Extracting Schema from Semi-structured Data

Svetlozar Nestorov

Serge Abiteboul

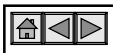
Rajeev Motwani

Stanford University

ACM SIGMOD Conference, 1998.

"Inferring Structure in Semi-structured Data"

Workshop on Management of Semi-structured Data, 1997.

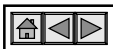


Outline

- **Introduction**
- **Lattice Approach**
- **Program Approach**
- **Experiment**
- **Conclusion**

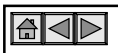
Introduction

- **Motivation**
  - homepages of group members
    - **name, e-mail, address, photo**
    - **similar information, but irregular schema**
  - perfect typing
    - **too large size**
    - **query optimization, graphical query interface**
- **Goal**
  - approximate typing
    - **tradeoff between the quality of a typing and its compactness**



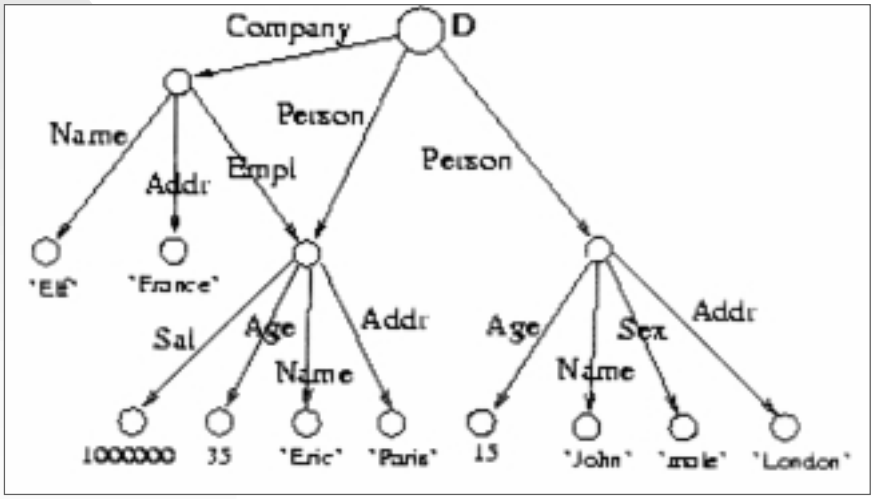
Introduction

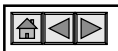
- **Preliminary**
  - data model
    - **Object Exchange Model (OEM)**
    - **labeled directed graph**
  - notation
    - **attribute(o)={labels on outgoing edges}**
    - **role(o)={labels on incoming edges}**
    - **at(S)=| {o| attribute(o)= S} |**
    - **above(S)=| {o| attribute(o)⊇S} |**
  - definition
    - **jump(S)=at(S)/above(S)**



Introduction

- **OEM Database D**



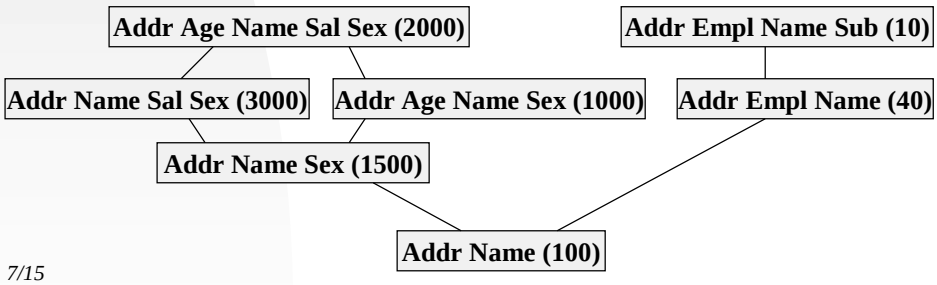


Lattice Approach

- **Algorithm**
  - identify candidate types
  - select types from candidates and organize them into a type hierarchy
  - derive the typing rules
  - validate the type hierarchy against the data
- **Counting lattice: at(S)**
  - significant jump: jump(S)≥θ

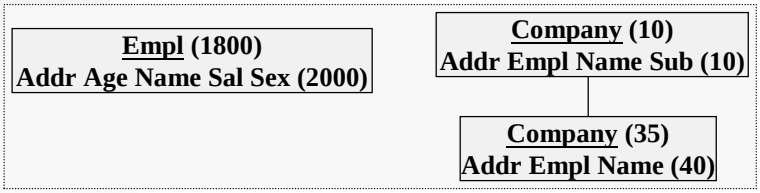
Lattice Approach

- **Counting lattice**
  - jump({Addr Age Name Sal Sex})=1
  - jump({Addr Empl Name Sub})=1
  - jump({Addr Empl Name})=0.8
  - jump({Addr Name Sal Sex})=0.6



Lattice Approach

- **Type hierarchy construction**
  - primary role: p-role(S)
    - ☞**the most frequent label in role(o) for all candidates**
  - select candidate T if there does not exist T' such that T'⊂T and p-role(T')=p-role(T)



Program Approach

- **Base relations**
- **Typing rule**
  - c(X):-A1&…&An
- **Type links**
  - link(Y,X,l)&c'(Y)
  - link(X,Y,l)&c'(Y)
  - link(X,Y,l)&atomic(Y,Z)
- **Example typing rule**
  - person(X):-link(X,Y,is-manager-of)& firm(Y)&link(X,Y',name)&atomic(Y',Z)
    - ☞**person(X):-is-manager-of^{firm},name^{atomic}**

link

FromObj	ToObj	Label
g	*m*	is-manager-of
j	*a*	is-manager-of
m	*g*	is-managed-by
a	*j*	is-managed-by
g	*gn*	name
j	*jn*	name
m	*mn*	name
a	*an*	name

atomic

Obj	Value
gn	"Gates"
jn	"Jobs"
mn	"Microsoft"
an	"Apple"

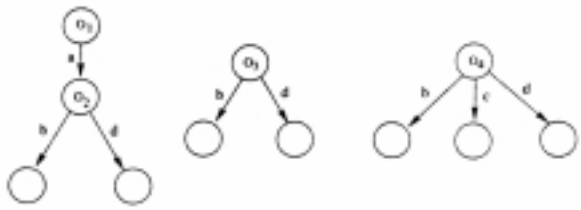
Program Approach

- **Example typing program**

$$
\begin{aligned}
\textit{project}: \quad \tau_1 &= \overleftarrow{Project}^3, \overleftarrow{Project}^4, \overleftarrow{Project}^5, \overrightarrow{Project\_Member}^3, \overrightarrow{Project\_Member}^4, \\
&\quad \overrightarrow{Name}^0, \overrightarrow{Homepage}^0 \\
\textit{publication}: \quad \tau_2 &= \overleftarrow{Publication}^3, \overleftarrow{Publication}^5, \overrightarrow{Author}^3, \overrightarrow{Name}^0, \overrightarrow{Conference}^0, \overrightarrow{Postscript}^0 \\
\textit{db-person}: \quad \tau_3 &= \overleftarrow{Project\_Member}^1, \overleftarrow{Group\_Member}^5, \overrightarrow{Project}^1, \overrightarrow{Birthday}^5, \overrightarrow{Degree}^6, \\
&\quad \overrightarrow{Years\_At\_Stanford}^0, \overrightarrow{Email}^0, \overrightarrow{Home\_Page}^0, \overrightarrow{Title}^0, \\
&\quad \overrightarrow{Name}^0, \overrightarrow{Original\_Home}^0, \overrightarrow{Personal\_Interest}^0, \overrightarrow{Research\_Interest}^0 \\
\textit{student}: \quad \tau_4 &= \overleftarrow{Project\_Member}^1, \overleftarrow{Student}^4, \overleftarrow{Group\_Member}^5, \overrightarrow{Project}^1, \overrightarrow{Advisor}^4, \\
&\quad \overrightarrow{Email}^0, \overrightarrow{Title}^0, \overrightarrow{Home\_Page}^0, \overrightarrow{Name}^0, \overrightarrow{Nickname}^0 \\
\textit{birthday}: \quad \tau_5 &= \overleftarrow{Birthday}^3, \overrightarrow{Namez}^0, \overrightarrow{Month}^0, \overrightarrow{Day}^0, \overrightarrow{Year}^0 \\
\textit{degree}: \quad \tau_6 &= \overleftarrow{Degree}^5, \overrightarrow{Major}^0, \overrightarrow{School}^0, \overrightarrow{Name}^0, \overrightarrow{Year}^0
\end{aligned}
$$

Program Approach

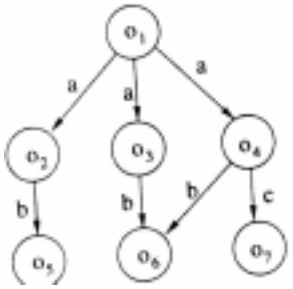
- **Defect**
  - excess
  - deficit
- **Example**



  - type1:-$\vec{a}^2$
  - type2:-$\overleftarrow{a}^1,\vec{b}^0,\vec{c}^0$
  - type3:-$\vec{b}^0,\vec{d}^0$
  - type assignment: o_4
    - **type2: excess=1, deficit=1**
    - **type3: excess=0, deficit=1**

Program Approach

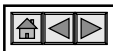
- **Minimal perfect typing**
  - step1
    - **type1:-$\vec{a}^2,\vec{a}^3,\vec{a}^4$**
    - **type2:-$\overleftarrow{a}^1,\vec{b}^0$**
    - **type3:-$\overleftarrow{a}^1,\vec{b}^0$**
    - **type4:- $\overleftarrow{a}^1,\vec{b}^0,\vec{c}^0$**



  - step2
    - **equivalent class: type2=type3**
  - step3
    - **type1:-$\vec{a}^2,\vec{a}^3$**
    - **type2:-$\overleftarrow{a}^1,\vec{b}^0$**
    - **type3:- $\overleftarrow{a}^1,\vec{b}^0,\vec{c}^0$**

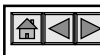
Program Approach

- **Clustering**
  - merge similar types
    - **type1:-$\vec{a}^0,\vec{b}^3$**
    - **type2:-$\vec{a}^0,\vec{b}^4$**
    - **type3:-$\vec{a}^0,\overleftarrow{b}^1$**
    - **type4:-$\vec{a}^0,\overleftarrow{b}^2$**
  - the order of coalescing has a significant effect on the quality of the results
  - distance function
    - **symmetric difference**
- **Recasting**

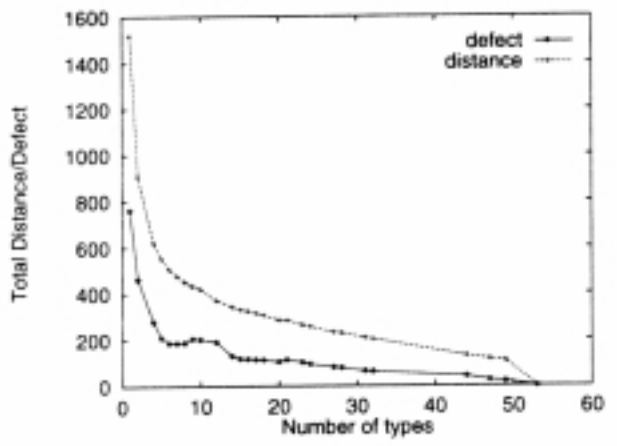


Experiment

Synthetic Data							*Typing*		
DB No	*Bipartite ?*	*Overlap ?*	*Perturb ?*	*Intended Types*	*Objects*	*Links*	*Perfect Types*	*Optimal Types*	*Defect*
1	Y	N	N	10	1500	2909	30	10	225
2	Y	N	Y	10	1500	2958	52	10	307
3	Y	Y	N	6	950	2409	19	6	239
4	Y	Y	Y	6	950	2442	35	6	283
5	N	N	N	5	400	726	317	5	181
6	N	N	Y	5	400	749	341	5	310
7	N	Y	N	5	400	775	375	5	291
8	N	Y	Y	5	400	795	381	5	333



Experiment



Conclusion

- **Contribution**
  - an algorithm for approximate typing of semi-structured data
  - each object may have more than one roles