Counting Types for Massive JSON Datasets

Mohamed-Amine Baazizi, Dario Colazzo,
Giorgio Ghelli, Carlo Sartiani

Counting types

- Can types count?
- Should they?

The problem

- Type inference for massive JSON datasets
- We infer this type

```
{       title :  Str ;
        text : [ Str ] + Null  ;
        author : { address:T? ; affil:T? ;… } ?
        abstract : Str ?
 }
```

- How «optional» is the author?

Let us count

$\{$ title : Str^{1000} ;

text : $([\ \mathrm{Str}^{8000}\ ]^{800} + \mathrm{Null}^{200})^{1000}$;

author : $\{$ add:T^{300} ?; affil:T^{300} ?;… $\}^{800}$?

abstract : Str^{20} ?

$\}^{1000}$

Correlation

- $\{\ \mathbf{addr{:}T}^{\mathbf{300}};\ \ \mathbf{aff{:}T}^{\mathbf{300}};\ r{:}T^{800}\ \}^{800}$
- $\{\ \mathbf{addr{:}T}^{\mathbf{300}};\ \ \mathbf{aff{:}T}^{\mathbf{300}};\ r{:}T^{300}\ \}^{300} + \{r{:}T^{500}\ \}^{500}$
- $\{\ \mathbf{addr{:}T}^{\mathbf{300}};\ r{:}T^{300}\ \}^{300} + \{\ \mathbf{aff{:}T}^{\mathbf{300}};\ r{:}T^{500}\ \}^{500}$
- $\{\ \mathbf{addr{:}T}^{\mathbf{300}};\ r{:}T^{500}\ \}^{500} + \{\ \mathbf{aff{:}T}^{\mathbf{300}};\ r{:}T^{300}\ \}^{300}$
- $\{\ \mathbf{addr{:}T}^{\mathbf{300}};\ r{:}T^{300}\ \}^{300} + \{\ \mathbf{aff{:}T}^{\mathbf{300}};\ r{:}T^{300}\ \}^{300} + \{r{:}T^{200}\}^{200}$

The type system

- $B ::= \text{Null}^i \mid \text{Int}^i \mid \text{Str}^i \mid \text{Bool}^i$
- $R ::= \{ l : T , \ldots , l : T \}^i$
- $A ::= [\,T\,]^i$
- $S ::= B \mid R \mid A$
- $T ::= S \mid 0 \mid T + T$
- $v : S$
- $v_1,\ldots,v_n \;:^{b}\; T$

The type inference algorithm

$$\frac{v_1 : S_1, \ldots, v_n : S_n}{\{l_1 : v_1, \ldots, l_n : v_n\} : \{l_1 : S_1, \ldots, l_n : S_n\}^{1}}$$

$$\frac{v_1, \ldots, v_n \ :^{\flat} T}{[v_1, \ldots, v_n] : [T]^{1}}$$

The type inference algorithm

- Multiset lossless rule:

$$\frac{v_1 : S_1 \quad \ldots \quad v_n : S_n}{v_1, \ldots, v_n :^{\flat} S_1 + \cdots + S_n}$$

- Merging rule:

$$\frac{v_1 : S_1 \quad \ldots \quad v_n : S_n}{v_1, \ldots, v_n :^{\flat} reduce(E)(S_1, \ldots, S_n)}$$

Parametric reduction

$$
\begin{aligned}
&\mathit{Reduce}(\mathbb{T}_1, \mathbb{T}_2, E) = \\
&\quad \bigoplus\Big(\quad \{\!|\, \mathit{Merge}(\mathbb{S}_1, \mathbb{S}_2, E) \mid \mathbb{S}_1 \in \circ\mathbb{T}_1, \mathbb{S}_2 \in \circ\mathbb{T}_2, E(\mathbb{S}_1, \mathbb{S}_2) \,|\!\}^{\mathrm{m}} \\
&\qquad\qquad \cup^{\mathrm{m}} \ \{\!|\ \mathbb{S}_1 \mid \mathbb{S}_1 \in \circ\mathbb{T}_1,\ \nexists \mathbb{S}_2 \in \circ\mathbb{T}_2.\ E(\mathbb{S}_1, \mathbb{S}_2)\ |\!\}^{\mathrm{m}} \\
&\qquad\qquad \cup^{\mathrm{m}} \ \{\!|\ \mathbb{S}_2 \mid \mathbb{S}_2 \in \circ\mathbb{T}_2,\ \nexists \mathbb{S}_1 \in \circ\mathbb{T}_1.\ E(\mathbb{S}_1, \mathbb{S}_2)\ |\!\}^{\mathrm{m}} \quad \Big)
\end{aligned}
$$

The type inference algorithm

Map

$$\frac{v_1 : T_1 \quad \dots \quad v_n : T_n}{v_1, \dots, v_n : reduce(E)(T_1, \dots, T_n)}$$

Combine/Reduce

The meaning of counting

- Nested counting: how?
  - [[2,2,2] , [2] , [2,[3,3],2] , [2,2]] , [[[3,3]]]
- Cumulative counting
  - $[\ \ [\ \mathrm{Int}^8 + [\mathrm{Int}^4\ ]^2\ \ ]^5\ ]^2$
- More precise subtypes
  - $[\ [\ \mathrm{Int}^8 + [\mathrm{Int}^2\ ]^1\ \ ]^4\ ]^1 + [\ [\ [\mathrm{Int}^2\ ]^1\ \ ]^1\ ]^1$
  - $[\ [\ \mathrm{Int}^6\ ]^3 + [\ \mathrm{Int}^2 + [\mathrm{Int}^2\ ]^1\ \ ]^1\ ]^1 + [\ [\ [\mathrm{Int}^2\ ]^1\ \ ]^1\ ]^1$
  - $[\ [\mathrm{Int}^3]^1 + [\mathrm{Int}^1]^1 + [\mathrm{Int}^2 + [\mathrm{Int}^2\ ]^1\ \ ]^1\ ] + [\mathrm{Int}^2]^1\ ]^1 + [[\ [\mathrm{Int}^2\ ]^1\ \ ]^1\ ]^1$

The meaning of counting

- The formal machinery to interpret counting types:
  - $⟦\text{ Int }⟧ = \{ \text{ Nat } \}$
  - $⟦\text{ Int }⟧^3 = \{ \{i, j, k\}^{b} \mid i \in \text{Nat}, j \in \text{Nat}, k \in \text{Nat}\}$

  - $⟦\, T + U \,⟧ = ⟦\, T \,⟧ \cup ⟦\, U \,⟧$
  - $⟦\, T + U \,⟧ = \{ M1 \cup M2 \mid (M1,M2) \text{ in } ⟦T⟧ \times ⟦U⟧ \}$
  - Hence T not a subtype of T+U

For example: twitter data

{ contributors: $(\mathrm{Null}^{9,599,980} + [\mathrm{Num}^{20}]^{20})^{9,600,000}$?;

retweeted : $\mathrm{Bool}^{9,600,000}$?;

retweeted_status {…} : $\{\ldots\}^{1,200,000}$?;

deleted : $\{\ldots\}^{300,000}$?;

$\}^{9,900,000}$

For example: twitter data

$\{ \text{con}: \ldots^{7{,}200{,}000}; \text{ret}: \text{Bool}^{7{,}200{,}000}; \ldots \}^{7{,}200{,}000}$

$+\{ \text{con}: \ldots^{1{,}200{,}000}; \text{ret}: \text{Bool}^{1{,}200{,}000}; \ldots \}^{1{,}200{,}000}$

$+\{ \text{con}: \ldots^{1{,}040{,}000}; \text{ret}: \text{Bool}^{1{,}040{,}000}; \text{r\_s}: \{\}^{1{,}040{,}000}; \ldots \}^{1{,}040{,}000}$

$+\{ \text{con}: \ldots^{160{,}000}; \text{ret}: \text{Bool}^{160{,}000}; \text{r\_s}: \{\}^{160{,}000}; \ldots \}^{160{,}000}$

$+\{ \text{deleted}: \{ \}^{300{,}000}; \ldots \}^{300{,}000}$:

To sum up

- An algorithm to summarize JSON data:
  - Easy
  - Well defined semantics
  - Parametric
  - Parallel
  - Yielding quantitative information
- Quantitative information from indexes to types
- What else may a counting type do?