DATABASE THEORY

Lecture 14: Datalog Implementation

David Carral
Knowledge-Based Systems

TU Dresden, May 26, 2020

Implementing Datalog

FO queries (and thus also COs and UCQs) are supported by almost all DBMS

How can Datalog queries be answered in practice?

Techniques for dealing with recursion in DBMS query answering

There are two major paradigms for answering recursive queries:

- **Bottom-up**: derive conclusions by applying rules to given facts
- **Top-down**: search for proofs to infer results given query

Review: Datalog

A rule-based recursive query language

```
father(alice, bob)
mother(alice, carla)
Parent(x, y) ← father(x, y)
Parent(x, y) ← mother(x, y)
SameGeneration(x, y)
SameGeneration(x, y) ← Parent(x, y) ∧ Parent(y, w) ∧ SameGeneration(y, w)
```

- Datalog is more complex than FO query answering
- Datalog is more expressive than FO query answering
- Semipositive Datalog with a successor ordering captures P
- Datalog containment is undecidable

Remaining question: How can Datalog query answering be implemented?

Computing Datalog Query Answers Bottom-Up

We already saw a way to compute Datalog answers bottom-up: the step-wise computation of the consequence operator \( T_F \)

Bottom-up computation is known under many names:

- **Forward-chaining** since rules are “chained” from premise to conclusion
  (common in logic programming)
- **Materialisation** since inferred facts are stored (“materialised”)
  (common in databases)
- **Saturation** since the input database is “saturated” with inferences
  (common in theorem proving)
- **Deductive closure** since we “close” the input under entailments
  (common in formal logic)
Naive Evaluation of Datalog Queries

A direct approach for computing $T_P^n$:

```
01 $T_P^0 := \emptyset$
02 $i := 0$
03 repeat:
04   $T_P^{i+1} := \emptyset$
05   for $H \leftarrow B_1 \land \ldots \land B_i$ in $P$:
06     for $\theta \in B_1 \land \ldots \land B_i(T_P^i)$:
07       $T_P^{i+1} := T_P^{i+1} \cup \{H(\theta)\}$
08     $i := i + 1$
09 until $T_P^{i+1} = T_P^i$
10 return $T_P^n$
```

Notation for line 06/07:
- a substitution $\theta$ is a mapping from variables to database elements
- for a formula $F$, we write $F\theta$ for the formula obtained by replacing each free variable $x$ in $F$ by $\theta(x)$
- for a CQ $Q$ and database $I$, we write $\theta \in Q(I)$ if $I \models Q\theta$

What’s Wrong with Naive Evaluation?

An example Datalog program:

- $\text{e}(1, 2)$ $\text{e}(2, 3)$ $\text{e}(3, 4)$ $\text{e}(4, 5)$
- $(R1)$ $T(x, y) \leftarrow \text{e}(x, y)$
- $(R2)$ $T(x, z) \leftarrow T(x, y) \land T(y, z)$

How many body matches do we need to iterate over?

- $T_P^0 = \emptyset$ initialisation
- $T_P^1 = \{T(1, 2), T(2, 3), T(3, 4), T(4, 5)\}$ 4 matches for $(R1)$
- $T_P^2 = T_P^1 \cup \{T(1, 3), T(2, 4), T(3, 5)\}$ 4 $\times (R1) + 3 \times (R2)$
- $T_P^3 = T_P^2 \cup \{T(1, 4), T(2, 5), T(1, 5)\}$ 4 $\times (R1) + 8 \times (R2)$
- $T_P^4 = T_P^3 \cup \{T(1, 6), \ldots\}$ 4 $\times (R1) + 10 \times (R2)$

In total, we considered 37 matches to derive 11 facts

Less Naive Evaluation Strategies

Does it really matter how often we consider a rule match?
After all, each fact is added only once . . .

In practice, finding applicable rules takes significant time, even if the conclusion does not need to be added – iteration takes time!
~ huge potential for optimisation

Observation:
we derive the same conclusions over and over again in each step

Idea: apply rules only to newly derived facts
~ semi-naive evaluation

Semi-Naive Evaluation

The computation yields sets $T_P^0 \subseteq T_P^1 \subseteq T_P^2 \subseteq \ldots \subseteq T_P^n$

- For an IDB predicate $R$, let $R'$ be the "predicate" that contains exactly the $R$-facts in $T_P^i$
- For $i \leq 1$, let $\Delta_R^i$ be the collection of facts $R' \setminus R^{i-1}$

We can restrict rules to use only some computations.
Some options for the computation in step $i + 1$:

- $T(x, z) \leftarrow T(x, y) \land T(y, z)$ same as original rule
- $T(x, z) \leftarrow \Delta_R^1(x, y) \land \Delta_R^1(y, z)$ restrict to new facts
- $T(x, z) \leftarrow \Delta_R^1(x, y) \land T(y, z)$ partially restrict to new facts
- $T(x, z) \leftarrow T(x, y) \land \Delta_R^1(y, z)$ partially restrict to new facts

What to choose?
Semi-Naive Evaluation (2)

Inferences that involve new and old facts are necessary:

\[ \text{e}(1, 2), \text{e}(2, 3), \text{e}(3, 4), \text{e}(4, 5) \]

(R1) \[ T(x, y) \leftarrow \text{e}(x, y) \]

(R2) \[ T(x, z) \leftarrow T(x, y) \land T(y, z) \]

\[ T_p^0 = \emptyset \]

\[ \Delta_1^T = \{T(1, 2), T(2, 3), T(3, 4), T(4, 5)\} \]

\[ T_p^1 = \Delta_1^T \]

\[ \Delta_2^T = \{T(1, 3), T(2, 4), T(3, 5)\} \]

\[ T_p^2 = T_p^1 \cup \Delta_2^T \]

\[ \Delta_3^T = \{T(1, 4), T(2, 5), T(1, 5)\} \]

\[ T_p^3 = T_p^2 \cup \Delta_3^T \]

\[ \Delta_4^T = \emptyset \]

\[ T_p^4 = T_p^3 \]

To derive \( T(1, 4) \) in \( \Delta_4^T \), we need to combine

\( T(1, 3) \in \Delta_3^T \) with \( T(3, 4) \in \Delta_4^T \) or \( T(1, 2) \in \Delta_2^T \) with \( T(2, 4) \in \Delta_4^T \).

\( \vdash \) rule \( T(x, z) \leftarrow \Delta_1^T(x, y) \land \Delta_4^T(y, z) \) is not enough

Semi-Naive Evaluation: Full Definition

**Correct approach:** consider only rule application that use at least one newly derived

IDB atom

For example program:

\[ \text{e}(1, 2), \text{e}(2, 3), \text{e}(3, 4), \text{e}(4, 5) \]

(R1) \[ T(x, y) \leftarrow \text{e}(x, y) \]

(R2.1) \[ T(x, z) \leftarrow \Delta_1^T(x, y) \land T(y, z) \]

(R2.2) \[ T(x, z) \leftarrow T^1(x, y) \land \Delta_4^T(y, z) \]

There is still redundancy here: the matches for \( T(x, z) \leftarrow \Delta_1^T(x, y) \land \Delta_4^T(y, z) \) are covered by both (R2.1) and (R2.2).

\( \vdash \) replace (R2.2) by the following rule:

(R2.2') \[ T(x, z) \leftarrow T^1(x, y) \land \Delta_4^T(y, z) \]

EDB atoms do not change, so their \( \Delta \) would be \( \emptyset \)

\( \vdash \) ignore such rules after the first iteration

Semi-Naive Evaluation: Example

\[ \text{e}(1, 2), \text{e}(2, 3), \text{e}(3, 4), \text{e}(4, 5) \]

(R1) \[ T(x, y) \leftarrow \text{e}(x, y) \]

(R2.1) \[ T(x, z) \leftarrow \Delta_1^T(x, y) \land T(y, z) \]

(R2.2') \[ T(x, z) \leftarrow T^1(x, y) \land \Delta_4^T(y, z) \]

How many body matches do we need to iterate over?

\[ T_p^0 = \emptyset \quad \text{initialisation} \]

\[ T_p^1 = \{T(1, 2), T(2, 3), T(3, 4), T(4, 5)\} \quad 4 \times (R1) \]

\[ T_p^2 = T_p^1 \cup \{T(1, 3), T(2, 4), T(3, 5)\} \quad 3 \times (R2.1) \]

\[ T_p^3 = T_p^2 \cup \{T(1, 4), T(2, 5), T(1, 5)\} \quad 3 \times (R2.1), 2 \times (R2.2') \]

\[ T_p^4 = T_p^3 \quad 1 \times (R2.1), 1 \times (R2.2') \]

In total, we considered 14 matches to derive 11 facts

Advantages and disadvantages:

- Huge improvement over naive evaluation
- Some redundant computations remain (see example)
- Some overhead for implementation (store level of entailments)
Summary and Outlook

Datalog queries can be evaluated bottom-up or top-down

Simplest practical bottom-up technique: semi-naive evaluation

Next question:
- Can we improve Datalog evaluation further?
- What about practical implementations?