Overview

1. Introduction | Relational data model
2. First-order queries
3. Complexity of query answering
4. Complexity of FO query answering
5. Conjunctive queries
6. Tree-like conjunctive queries
7. Query optimisation
8. Conjunctive Query Optimisation / First-Order Expressiveness
9. First-Order Expressiveness / Introduction to Datalog
10. Expressive Power and Complexity of Datalog
11. Optimisation and Evaluation of Datalog
12. Evaluation of Datalog (2)
13. Graph Databases and Path Queries
14. Outlook: database theory in practice

See course homepage [⇒ link] for more information and materials

Markus Krötzsch, 16 June 2016 Database Theory slide 2 of 29

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, x) |
| SameGeneration(x, y) ← Parent(x, v) ∧ Parent(y, w) ∧ SameGeneration(v, w) |

There are three equivalent ways of defining Datalog semantics:

- Proof-theoretic: What can be proven deductively?
- Operational: What can be computed bottom up?
- Model-theoretic: What is true in the least model?

Next questions:

- What can we express in this language?
- How hard is it in terms of complexity?

Datalog and UCQs

We have seen in the exercise that UCQs can be expressed in Datalog. Let's make this relationship more precise.

For a Datalog program $P$:

- An IDB predicate $R$ depends on an IDB predicate $S$ if $P$ contains a rule with $R$ in the head and $S$ in the body.
- $P$ is non-recursive if there is no cyclic dependency.

**Theorem**

UCQs have the same expressivity as non-recursive Datalog.

That is: a query mapping can be expressed by some UCQ if and only if it can be expressed by a non-recursive Datalog program.

However, Datalog can be exponentially more succinct (shorter queries), as illustrated in exercise.
Datalog and Domain Independence

Domain independence was considered useful for FO queries → results should not change if domain changes

Several solutions:
- **Active domain semantics**: restrict to elements mentioned in database or query
- **Domain-independent queries**: restrict to query where domain does not matter
- **Safe-range queries**: decidable special case of domain independence

Our definition of Datalog uses the active domain (=Herbrand universe) to ensure domain independence

Markus Krötzsch, 16 June 2016

Safe Datalog Queries

Similar to safe-range FO queries, there are also simple syntactic conditions that ensure domain independence for Datalog:

**Definition**

A Datalog rule is **safe** if all variables in its head also occur in its body. A Datalog program/query is safe if all of its rules are.

Simple observations:
- safe Datalog queries are domain independent
- every Datalog query can be expressed as a safe Datalog query ...
- ... and un-safe queries are not much more succinct either (exercise)

Some texts require Datalog queries to be safe in general but in most contexts there is no real need for this

Markus Krötzsch, 16 June 2016

Complexity of Datalog

How hard is answering Datalog queries?

Recall:
- Combined complexity: based on query and database
- Data complexity: based on database; query fixed
- Query complexity: based on query; database fixed

Plan:
- First show upper bounds (outline efficient algorithm)
- Then establish matching lower bounds (reduce hard problems)

Markus Krötzsch, 16 June 2016

A Simpler Problem: Ground Programs

Let’s start with Datalog without variables → sets of ground rules a.k.a. propositional Horn logic program

**Naive computation of** $T_P^\infty$:

1. $T_p^0 := \emptyset$
2. $i := 0$
3. **repeat**:
4. $T_p^{i+1} := \emptyset$
5. **for** $H \leftarrow B_1 \land \ldots \land B_i \in P$:
6. **if** $\{B_1, \ldots, B_i\} \subseteq T_p^i$:
7. $T_p^{i+1} := T_p^{i+1} \cup \{H\}$
8. $i := i + 1$
9. **until** $T_p^{i+1} = T_p^{i}$
10. **return** $T_p$

How long does this take?
- At most $|P|$ facts can be derived
- Algorithm terminates with $i \leq |P| + 1$
- In each iteration, we check each rule once (linear), and compare its body to $T_p$ (quadratic)

$\leadsto$ polynomial runtime
Complexity of Propositional Horn Logic

Much better algorithms exist:

**Theorem (Dowling & Gallier, 1984)**

For a propositional Horn logic program $P$, the set $T^\infty_P$ can be computed in linear time.

Nevertheless, the problem is not trivial:

**Theorem**

For a propositional Horn logic program $P$ and a proposition (or ground atom) $A$, deciding if $A \in T^\infty_P$ is a $P$-complete problem.

Remark:

all $P$ problems can be reduced to propositional Horn logic entailment yet not all problems in $P$ (or even in $NL$) can be solved in linear time!

Markus Krötzsch, 16 June 2016

Datalog Complexity: Upper Bounds

A straightforward approach:

1. Compute the grounding $ground(P)$ of $P$ w.r.t. the database $I$
2. Compute $T^\infty_{ground(P)}$

Complexity estimation:

- The number of constants $N$ for grounding is linear in $P$ and $I$
- A rule with $m$ distinct variables has $N^m$ ground instances
- Step (1) creates at most $|P| \cdot N^M$ ground rules, where $M$ is the maximal number of variables in any rule in $P$
  - $ground(P)$ is polynomial in the size of $I$
  - $ground(P)$ is exponential in $P$
- Step (2) can be executed in linear time in the size of $ground(P)$

Summing up: the algorithm runs in $P$ data complexity and in $ExpTime$ query and combined complexity

Markus Krötzsch, 16 June 2016

P-Hardness of Data Complexity

These upper bounds are tight:

**Theorem**

Datalog query answering is:

- $ExpTime$-complete for combined complexity
- $ExpTime$-complete for query complexity
- $P$-complete for data complexity

It remains to show the lower bounds.

Markus Krötzsch, 16 June 2016
Preparing for a Long Computation

We need to encode $2^{N_k}$ time points and tape positions. This can be achieved by using binary numbers with $N_k$ digits.

Easy observations:
- True$(A)$ is derived if and only if $A$ is a consequence of the original propositional program.
- The encoding of propositional programs as databases can be computed in logarithmic space.
- The Datalog program is the same for all propositional programs.

$\leadsto$ Datalog query answering is P-hard for data complexity.

Defining a Long Chain

We can define succ$^i(X, Y)$, first$^i(X)$, and last$^i(X)$ as follows:

\begin{align*}
\text{succ}^1(0, 1) & \quad \text{first}^1(0) \quad \text{last}^1(1) \\
\text{succ}^{i+1}(0, X, 0, Y) & \leftarrow \text{succ}^i(X, Y) \\
\text{succ}^{i+1}(1, X, 1, Y) & \leftarrow \text{succ}^i(X, Y) \\
\text{succ}^{i+1}(0, X, 1, Y) & \leftarrow \text{last}^i(X) \land \text{first}^i(Y) \\
\text{first}^{i+1}(0, X) & \leftarrow \text{first}^i(X) \\
\text{last}^{i+1}(1, X) & \leftarrow \text{last}^i(X)
\end{align*}

For $X = x_1, \ldots, x_i$ and $Y = y_1, \ldots, y_i$, and the arity of head is $2 \cdot N_k$.

Now for $M = N_k$, we define $\leq^M(X, Y)$ as the reflexive, transitive closure of succ$^M(X, Y)$:

\begin{align*}
\leq^M(X, X) & \leftarrow \\
\leq^M(X, Z) & \leftarrow \leq^M(X, Y) \land \text{succ}^M(Y, Z)
\end{align*}

ExpTime-Hardness of Query Complexity

A direct proof:

Encode the computation of a deterministic Turing machine for up to exponentially many steps.

Recall that ExpTime = $\bigcup_{k \geq 1} \text{TIME}(2^k)$

- In our case, $n = N$ is the number of database constants.
- $k$ is some constant.

$\leadsto$ we need to simulate up to $2^{N_k}$ steps (and tape cells).

Main ingredients of the encoding:
- state$_q(X)$: the TM is in state $q$ after $X$ steps.
- head$(X, Y)$: the TM head is at tape position $Y$ after $X$ steps.
- symbol$_\sigma(X, Y)$: the tape cell at position $Y$ holds symbol $\sigma$ after $X$ steps.

$\leadsto$ How to encode $2^{N_k}$ time points $X$ and tape positions $Y$?
Initialising the Computation

We can now encode the initial configuration of the Turing Machine for an input word \( \sigma_1 \cdots \sigma_n \in (\Sigma \setminus \{\lambda\})^* \).

We write \( B_i \) for the binary encoding of a number \( i \) with \( M = \lambda^k \) digits, and \( Y = y_1, \ldots, y_M \).

\[
\begin{align*}
\text{state}_{q_0}(B_0) & \quad \text{where } q_0 \text{ is the TM's initial state} \\
\text{head}(B_0, B_0) \\
\text{symbol}_{\sigma_1}(B_0, B_i) & \quad \text{for all } i \in \{1, \ldots, n\} \\
\text{symbol}_{\sigma}(B_0, Y) & \leftarrow \leq^M(B_{n+1}, Y)
\end{align*}
\]

Hardness Results

**Lemma**

A deterministic TM accepts an input in \( \text{TIME}(2^k) \) if and only if the Datalog program defined above entails the fact \( \text{accept}(()) \).

We obtain \( \text{ExpTime} \)-hardness of Datalog query answering:

- The decision problem of any language in \( \text{ExpTime} \) can be solved by a deterministic TM in \( \text{TIME}(2^k) \) for some constant \( k \)
- In particular, there are \( \text{ExpTime} \)-hard languages \( L \) with suitable deterministic TM \( M \) and constant \( k \)
- For any input word \( w \), we can reduce acceptance of \( w \) by \( M \) in \( \text{TIME}(2^k) \) to entailment of \( \text{accept}(()) \) by a Datalog program \( P(w, M, k) \)
- \( P(w, M, k) \) is polynomial in \( k \) and the size of \( M \) and \( w \) (in fact, it can be constructed in logarithmic space)

TM Transition and Acceptance Rules

For each transition \( \langle q, \sigma, q', \sigma', d \rangle \in \Delta \), we add rules:

\[
\begin{align*}
\text{symbol}_{\sigma'}(X', Y) & \leftarrow \text{succ}^M(X, X') \land \text{head}(X, Y) \land \text{symbol}_{\sigma}(X, Y) \land \text{state}_q(X) \\
\text{state}_q(X') & \leftarrow \text{succ}^M(X, X') \land \text{head}(X, Y) \land \text{symbol}_{\sigma}(X, Y) \land \text{state}_q(X)
\end{align*}
\]

Similar rules are used for inferring the new head position (depending on \( d \))

Further rules ensure the preservation of unaltered tape cells:

\[
\begin{align*}
\text{symbol}_{\sigma}(X', Y) & \leftarrow \text{succ}^M(X, X') \land \text{symbol}_{\sigma}(X, Y) \land \text{head}(X, Z) \land \text{succ}^M(Z, Z') \land \leq^M(Z', Y) \\
\text{symbol}_{\sigma}(X', Y) & \leftarrow \text{succ}^M(X, X') \land \text{symbol}_{\sigma}(X, Y) \land \text{head}(X, Z) \land \text{succ}^M(Z', Z) \land \leq^M(Y, Z')
\end{align*}
\]

The TM accepts if it ever reaches the accepting state \( q_{\text{acc}} \):

\[
\text{accept()} \leftarrow \text{state}_{q_{\text{acc}}}(X)
\]

ExpTime-Hardness: Notes

Some further remarks on our construction:

- The constructed program does not use EDB predicates \( \rightarrow \) database can be empty
- Therefore, hardness extends to query complexity
- Using a fixed (very small) database, we could have avoided the use of constants
- We used IDB predicates of unbounded arity \( \rightarrow \) they are essential for the claimed hardness
The Big Picture

Where does Datalog fit in this picture?

Datalog Expressivity and Homomorphisms

How can we know that something is not expressible in Datalog?

A useful property: Datalog is "closed under homomorphisms"

**Theorem**

Consider a Datalog program $P$, an atom $A$, and databases $I$ and $J$. If $P$ entails $A$ over $I$, and there is a homomorphism $\mu$ from $I$ to $J$, then $\mu(P)$ entails $\mu(A)$ over $J$.

(By $\mu(P)$ and $\mu(A)$ we mean the program/atom obtained by replacing constants in $P$ and $A$, respectively, by their $\mu$-images.)

**Proof (sketch):**

- Closure under homomorphism holds for conjunctive queries
- Single rule applications are like conjunctive queries
- We can show the claim for all $T_{P,I}^i$ by induction on $i$

Limits of Datalog Expressiveness

Closure under homomorphism shows many limits of Datalog

Special case: there is a homomorphism from $I$ to $J$ if $I \subseteq J$

$\Rightarrow$ Datalog entailments always remain true when adding more facts

- **Parity** can not be expressed
- **Connectivity** can not be expressed
- It cannot be checked if the input database is a chain
- ... 

However this criterion is not sufficient!

Datalog cannot even express all polynomial time query mappings that are closed under homomorphism

Expressivity of Datalog

Datalog is $P$-complete for data complexity:

- Entailments can be computed in polynomial time with respect to the size of the input database $I$
- There is a Datalog program $P$, such that all problems that can be solved in polynomial time can be reduced to the question whether $P$ entails some fact over a database $I$ that can be computed in logarithmic space.

$\Rightarrow$ So Datalog can solve all polynomial problems?

No, it can't. Many problems in $P$ that cannot be solved in Datalog:

- **Parity**: Is the number of elements in the database even?
- **Connectivity**: Is the input database a connected graph?
- Is the input database a chain (or linear order)?
- ...
Capturing \( \text{PTime} \) in Datalog

How could we extend Datalog to capture all query mappings in \( \text{P} \)?

\( \rightarrow \) semipositive Datalog on an ordered domain

**Definition**

Semipositive Datalog, denoted Datalog\(^+\), extends Datalog by allowing negated EDB atoms in rule bodies. Datalog (semipositive or not) with a successor ordering assumes that there are special EDB predicates succ (binary), first and last (unary) that characterise a total order on the active domain.

Semipositive Datalog with a total order corresponds to standard Datalog on extended databases:

- For each ground fact \( r(c_1, \ldots, c_n) \) with \( I \not\models r(c_1, \ldots, c_n) \), add a new fact \( \overline{r}(c_1, \ldots, c_n) \) to \( I \), using a new EDB predicate \( \overline{r} \)
- Replace all uses of \( \neg r(t_1, \ldots, t_n) \) in \( P \) by \( \overline{r}(t_1, \ldots, t_n) \)
- Define extensions for the EDB predicates succ, first and last to characterise some (arbitrary) total order on the active domain.

Datalog Expressivity: Summary

The \( \text{PTime} \) capturing result is a powerful and exhaustive characterisation for semipositive Datalog with a successor ordering.

Situation much less clear for other variants of Datalog (as of 2015):

- What exactly can we express in Datalog without EDB negation and/or successor ordering?
  - Does a weaker language suffice to capture \( \text{PTime} \)? \( \rightarrow \) No!
  - When omitting negation, do we get query mappings closed under homomorphism? No!\(^1\)

- How about query mappings in \( \text{PTime} \) that are closed under homomorphism?
  - Does plain Datalog capture these? \( \rightarrow \) No!\(^2\)
  - Does Datalog with successor ordering capture these? \( \rightarrow \) No!\(^3\)

\(^1\)Counterexample on previous slide
\(^2\)[A. Dawar, S. Kreutzer, ICALP 2008]
\(^3\)[S. Rudolph, M. Thomazo, IJCAI 2016]
Summary and Outlook

Non-recursive Datalog can express UCQs

Datalog is more complex than FO query answering:
- $\text{ExpTime}$-complete for query and combined complexity
- $\text{P}$-complete for data complexity

Datalog cannot express all query mappings in $\text{P}$ but semipositive Datalog with a successor ordering can

Next topics:
- Query containment for Datalog
- Implementation techniques for Datalog