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, 28 April 2016

Review: FO Query Complexity

The evaluation of FO queries is
- PSPACE-complete for combined complexity
- PSPACE-complete for query complexity
- AC\(^0\)-complete for data complexity

~ PSPACE is rather high
~ Are there relevant query languages that are simpler than that?

Markus Krötzsch, 28 April 2016

Conjunctive Queries

Idea: restrict FO queries to conjunctive, positive features

Definition

A conjunctive query (CQ) is an expression of the form

\[ \exists y_1, \ldots, y_m. A_1 \land \ldots \land A_t \]

where each \( A_i \) is an atom of the form \( R(t_1, \ldots, t_k) \). In other words, a conjunctive query is an FO query that only uses conjunctions of atoms and (outer) existential quantifiers.

Example: “Find all lines that depart from an accessible stop” (as seen in earlier lectures)

\[ \exists y_{\text{SID}}, y_{\text{Stop}}, y_{\text{To}}. \text{Stops}(y_{\text{SID}}, y_{\text{Stop}}, \text{"true"}) \land \text{Connect}(y_{\text{SID}}, y_{\text{To}}, x_{\text{Line}}) \]
Conjunctive Queries in Relational Calculus

The expressive power of CQs can also be captured in the relational calculus.

**Definition**

A conjunctive query (CQ) is a relational algebra expression that uses only the operations select $\sigma_{n=m}$, project $\pi_{a_1,\ldots,a_n}$, join $\bowtie$, and renaming $\delta_{a_1,\ldots,a_n \rightarrow b_1,\ldots,b_n}$.

Renaming is only relevant in named perspective.

CQs are also known as SELECT-PROJECT-JOIN queries.

---

Extensions of Conjunctive Queries

Two features are often added:

- **Equality**: CQs with equality can use atoms of the form $t_1 \approx t_2$ (in relational calculus: table constants)
- **Unions**: unions of conjunctive queries are called UCQs (in this case the union is only allowed as outermost operator)

Both extensions truly increase expressive power (as shown in exercise).

Features omitted on purpose: negation and universal quantifiers

The reason for this is query complexity (as we shall see).

---

Boolean Conjunctive Queries

A Boolean conjunctive query (BCQ) asks for a mapping from query variables to domain elements such that all atoms are true.

**Example**: “Is there an accessible stop where some line departs?”

$$\exists y_{\text{SID}}, y_{\text{Stop}}, y_{\text{To}}, y_{\text{Line}}. \text{Stops}(y_{\text{SID}}, y_{\text{Stop}}, \text{"true"}) \land \text{Connect}(y_{\text{SID}}, y_{\text{To}}, y_{\text{Line}})$$

---

How Hard is it to Answer CQs?

If we know the variable mappings, it is easy to check:

- Checking if a single ground atom $R(c_1, \ldots, c_k)$ holds can be done in linear time.
- Checking if a conjunction of ground atoms holds can be done in quadratic time.

A candidate BCQ match can be verified in $\mathbf{P}$.

(There are $n^m$ candidates: $n$ size of domain; $m$ number of query variables)

**Theorem**

BCQ query answering is in $\mathbf{NP}$ for combined complexity (and also for query complexity).

Better than $\mathbf{PSPACE}$ (presumably).
Can we do any better?

Not really. To see this, let’s look at some other problems.

Consider two relational structures $I$ and $J$ (= database instances, interpretations, hypergraphs)

**Definition**

A homomorphism $h$ from $I$ to $J$ is a function $h : \Delta^I \rightarrow \Delta^J$ such that, for all relation names $R$:

$$\text{if } \langle d_1, \ldots, d_n \rangle \in R^I \text{ then } \langle h(d_1), \ldots, h(d_n) \rangle \in R^J.$$

The homomorphism problem is the question if there is a homomorphism from $I$ to $J$.

**Example: Three-colouring as Homomorphism**

$I$:

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
</tr>
</tbody>
</table>

$J$:

3-colouring is NP-hard

$\leadsto$ the homomorphism problem is NP-hard

**Complexity of Conjunctive Query Answering**

We showed that BCQ answering is in NP and that the homomorphism problem is NP-hard, therefore:

**Theorem**

BCQ answering is

- NP-complete for combined complexity
- NP-complete for query complexity
- in $AC^0$ for data complexity (inherited from FO queries)
Constraint Satisfaction Problems

Another important problem equivalent to BCQ answering

Definition

A constraint satisfaction problem (CSP) over a domain $\Delta$ is given by a set of variables $\{x_1, \ldots, x_n\}$ and a set of constraints $\{C_1, \ldots, C_m\}$, where each constraint $C_i$ has the form $\langle X_i, R_i \rangle$ with

- $X_i$ a list of variables from $\{x_1, \ldots, x_n\}$,
- $R_i$ a $|X_i|$-ary relation over $\Delta$.

A solution to the CSP is an assignment of variables to values from $\Delta$ such that all constraints are satisfied (=all tuples occur in the respective relations).

$\leadsto$ alternative notation for BCQ answering/homomorphism problem

Equivalent Problems

Summing up, the following problems are equivalent:

- Answering a conjunctive query over a database instance
- Finding a homomorphism from a relational structure to another
- Solving a constraint satisfaction problem

Each of these problems is NP-complete

Towards Better Complexities

NP-complete problems are still intractable

$\leadsto$ can we do better?

Problem: searching a match may require backtracking, eventually exploring all options

Intuition: life would be easier if we would not have to go back so much . . .

$\leadsto$ the problem is with the cycles

CSP Example

A combinatorial crossword puzzle:

Domain: $\Delta = \{A, \ldots, Z\}$

Variables: $x_1, \ldots, x_{26}$

Constraints:

5 vertically:

<table>
<thead>
<tr>
<th>1 horizontally:</th>
</tr>
</thead>
<tbody>
<tr>
<td>H E A R T</td>
</tr>
<tr>
<td>H O N E Y</td>
</tr>
<tr>
<td>I R O N Y</td>
</tr>
<tr>
<td>L A B O R</td>
</tr>
<tr>
<td>L A T E R</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

1 vertically:

<table>
<thead>
<tr>
<th>1 horizontally:</th>
</tr>
</thead>
<tbody>
<tr>
<td>B A C K D O</td>
</tr>
<tr>
<td>R E T R O</td>
</tr>
<tr>
<td>Y A C H T</td>
</tr>
<tr>
<td>Y E R B A</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Markus Krötzsch, 28 April 2016
Example: Cyclic CQs

"Is there a child whose parents are married with each other?"

\[ \exists y_c, y_m, y_f . \text{mother}(y_c, y_m) \land \text{father}(y_c, y_f) \land \text{married}(y_m, y_f) \]

\[ \leadsto \text{cyclic query} \]

Example: Acyclic CQs

"Is there a child whose parents are married with someone?"

\[ \exists y_c, y_m, y_f, y_{mm}, y_{mf} . \text{mother}(y_c, y_m) \land \text{father}(y_c, y_f) \land \text{married}(y_m, y_{mm}) \land \text{married}(y_{mf}, y_f) \]

\[ \leadsto \text{acyclic query} \]

Defining Acyclic Queries

Queries in general are hypergraphs

\[ \leadsto \text{What does “acyclic” mean?} \]

View hypergraphs as graphs to check acyclicity?

- **Primal graph**: same vertices; edges between each pair of vertices that occur together in a hyperedge
- **Incidence graph**: vertices and hyperedges as vertices, with edges to mark incidence (bipartite graph)

However: both graphs have cycles in almost all cases

Acyclic Hypergraphs

GYO-reduction algorithm to check acyclicity:

(1) Delete all vertices that occur in at most one hyperedge
(2) Delete all hyperedges that are empty or that are contained in other hyperedges

**Definition**

A hypergraph is **acyclic** if its GYO-reduct is \( \langle \emptyset, \emptyset \rangle \).
A CQ is **acyclic** if its associated hypergraph is.
Example 1: GYO-Reduction

Example 2: GYO-Reduction

Alternative Version of GYO-Reduction

An ear of a hypergraph \( (V, E) \) is a hyperedge \( e \in E \) that satisfies one of the following:

(1) there is an edge \( e' \in E \) such that \( e \neq e' \) and every vertex of \( e \) is either only in \( e \) or also in \( e' \), or

(2) \( e \) has no intersection with any other hyperedge.

Example:

\[ \sim \text{ edges } (4, 5, 6) \text{ and } (7, 8, 9) \text{ are ears} \]
GYO'-Reduction

Input: hypergraph $H = \langle V, E \rangle$
Output: GYO'-reduct of $H$

Apply the following simplification rule as long as possible:
- Select an ear $e$ of $H$
- Delete $e$
- Delete all vertices that only occurred in $e$

The GYO-reduct is $\langle \emptyset, \emptyset \rangle$ if and only if the GYO'-reduct is $\langle \emptyset, \emptyset \rangle$

Join Trees

Both GYO algorithms can be implemented in linear time

Open question: what benefit does BCQ acyclicity give us?

Fact: if a BCQ is acyclic, then it has a join tree

A join tree of a (B)CQ is an arrangement of its query atoms in a tree structure $T$, such that for each variable $x$, the atoms that refer to $x$ are a connected subtree of $T$.

A (B)CQ that has a join tree is called a tree query.

Processing Join Trees Efficiently

Join trees can be processed in polynomial time

Key ingredient: the semijoin operation

Given two relations $R[U]$ and $S[V]$, the semijoin $R^I \bowtie S^I$ is defined as $\pi_U(R^I \bowtie \lefttriangle S^I)$.

Join trees can now be processed by computing semijoins bottom-up

$\leadsto$ Yannakakis' Algorithm
Yannakakis' Algorithm by Example

Yannakakis' Algorithm: Summary

Polynomial time procedure for answering BCQs

Does not immediately compute answers in the version given here
\( \implies \) modifications needed

Even tree queries can have exponentially many results, but each can be computed (not just checked) in \( P \)
\( \implies \) output-polynomial computation of results

Summary and Outlook

Conjunctive queries (CQs) are an important special case of FO queries

Boolean CQ answering, the homomorphism problem and constraint satisfaction problems are equivalent and \( NP \)-complete

CQ answering is simpler, namely in \( P \), when CQs are tree queries

- Check acyclicity with GYO algorithm
- Evaluate query using Yannakakis' Algorithm

Open questions:

- Tree queries are rather special. Are there more general conditions for good queries?
- What about query optimisation?