Databases

Conjunctive Queries

Pierre Senellart

14 February 2024

Data complexity

Theorem

CQ evaluation is AC^0 in data complexity.

Proof.

CQs are a particular case of relational calculus queries. □

Combined complexity

Theorem

CQ evaluation is NP-complete in combined complexity.

Proof.

Membership in NP is easy. Hardness for NP can be proved by reduction from graph 3-colorability. □

α-acyclic query

- A CQ can be seen as a hypergraph (vertices are variables, hyperedges the atoms of the CQ, labeled by the relation name)
- A hypergraph $\mathcal{H}$ has a join tree where one can find a tree whose nodes are labeled by the hyperedges of $\mathcal{H}$ and such that:
  - every hyperedge of $\mathcal{H}$ appears as the label of one node of the tree;
  - for every vertex x of $\mathcal{H}$, the set of tree nodes labeled by a hyperedge referring to x is a connected subtree
- A query is α-acyclic if its hypergraph has a join tree
- Can be obtained in linear time if it exists [Tarjan and Yannakakis, 1984]

Yannakakis's algorithm [Yannakakis, 1981]

- Algorithm to evaluate acyclic queries (non-necessarily Boolean):
  1. Construct the join tree
  2. Eliminate all useless tuples of a relation with the semijoin operator $\ltimes$: $R \ltimes S = \Pi_{R.*}(R \bowtie S)$ by navigating twice in the join tree: from bottom up, then from top down
  3. Evaluate the query bottom up, by computing joins following the tree and by projecting useless variables out as you go
- Polynomial complexity in the size of the query, the input, and the output (combined complexity)

Case of the conjunctive queries

- We consider conjunctive queries (CQ) of the form:

$$q(\mathbf{x}) \leftarrow \exists \mathbf{y} : R_1(\mathbf{z}_1) \wedge \cdots \wedge R_n(\mathbf{z}_n)$$

  where each $\mathbf{z}_i$ is a tuple of variables among $\mathbf{x}$ and $\mathbf{y}$, and where each x_j appears at least in one $\mathbf{z}_j$
- Set semantics: for all database D, $q(D)$ is a finite set of tuples

Homomorphism

Definition

A homomorphism from a CQ q to a CQ q' is a function φ from the variables $\mathbf{x}, \mathbf{y}$ of q fo the variables $\mathbf{x}', \mathbf{y}'$ of q' such that:

- $\varphi(\mathbf{x}) = \mathbf{x}'$
- for every atom $R(\mathbf{z}_i)$ of q, there exists an atom $R(\mathbf{z}'_{i'})$ of q' such that $\varphi(\mathbf{z}_i) = \mathbf{z}'_{i'}$

Definition

A homomorphism is an isomorphism if it is one-to-one and its converse is a homomorphism.

Instance associated to a query

Definition

For all conjunctive query

$$q(\mathbf{x}) \leftarrow \exists \mathbf{y} : R_1(\mathbf{z}_1) \wedge \cdots \wedge R_n(\mathbf{z}_n)$$

one can construct the instance associated to q, denoted I_q, where the active domain is $\{ a_z \mid z \in \mathbf{x} \cup \mathbf{y} \}$ and which is formed of the n tuples $R(a_{z_{i1},\ldots,z_{ik}})$ for $R(z_{i1}, \ldots, z_{ik})$ atom of q

Instance associated to a query

Definition

For all conjunctive query

$$q(\mathbf{x}) \leftarrow \exists \mathbf{y} : R_1(\mathbf{z}_1) \wedge \cdots \wedge R_n(\mathbf{z}_n)$$

one can construct the instance associated to q, denoted I_q, where the active domain is $\{ a_z \mid z \in \mathbf{x} \cup \mathbf{y} \}$ and which is formed of the n tuples $R(a_{z_{i1},\ldots,z_{ik}})$ for $R(z_{i1}, \ldots, z_{ik})$ atom of q

Proposition

For all CQs $q(\mathbf{x})$, $q'(\mathbf{x}')$, there exists a homomorphism from q to q' iff $(a_{x'_1}, \ldots, a_{x'_j}) \in q(I_{q'})$.

Homomorphism theorem

Theorem ([Chandra and Merlin, 1977])

For all CQs q, q', $q \sqsubseteq q'$ iff there exists a homomorphism from q' to q.

Minimal query

Definition

A conjunctive query is minimal if it has a minimal number of atoms among all equivalent conjunctive queries.

Minimal query

Definition

A conjunctive query is minimal if it has a minimal number of atoms among all equivalent conjunctive queries.

- Translation of a CQ to an algebra query: if there are n atoms, we obtain $n - 1$ joins
- Joins are the most costly operations of the relational algebra (bar cross products)
- Finding a minimal query amounts to global optimization

Unicity of minimal query

Proposition ([Chandra and Merlin, 1977])

Let q be a CQ. Then there exists a CQ q' obtained by removing atoms from q which is minimal.

Proof.

Consider a minimal query equivalent to q and apply the homomorphism theorem. □

Proposition ([Chandra and Merlin, 1977])

Let q, q' be two equivalent minimal CQs. Then there exists an isomorphism from q to q'.

Proof.

Apply the homomorphism theorem. The image by the homomorphism is an equivalent minimal query. □

Minimization algorithm

Apply the following procedure to minimize a query:

> *For every atom of the query, test if there exists an equivalent query not containing this atom, and thus if there exists a homomorphism sending this atom to another atom of the query. If so, delete it, and continue until obtaining an equivalent minimal query.*

Complexity issues

Proposition

The following problems are NP-complete:

- given two CQs q, q', determine whether $q \sqsubseteq q'$
- given two CQs q, q', determine whether $q \equiv q'$
- given a CQ q, determine if q is non-minimal

Proof.

NP-hardness is by reduction from 3-colorability, as for combined complexity of query evaluation. Membership in NP is direct. □

Complexity issues

Proposition

The following problems are NP-complete:

- given two CQs q, q', determine whether $q \sqsubseteq q'$
- given two CQs q, q', determine whether $q \equiv q'$
- given a CQ q, determine if q is non-minimal

Proof.

NP-hardness is by reduction from 3-colorability, as for combined complexity of query evaluation. Membership in NP is direct. □

NP-hard... in the queries. Queries may be small enough so that an exponential algorithm may not be an issue.

Bag semantics [Chaudhuri and Vardi, 1993]

- In practice, RDBMSs implement a bag semantics
- Two queries in bag semantics are equivalent if and only if they are isomorphic (intuitively, because two similar but non isomorphic queries can introduce a different number of results)
- Query containment: Π_2^{P}-hard claimed (but not proved). Decidability (and precise complexity if decidable): open!

Plan

Bibliography I

Ashok K. Chandra and Philip M. Merlin. Optimal implementation of conjunctive queries in relational data bases. In *Proceedings of the 9th Annual ACM Symposium on Theory of Computing, May 4-6, 1977, Boulder, Colorado, USA*, pages 77–90, 1977. doi: 10.1145/800105.803397. URL `http://doi.acm.org/10.1145/800105.803397`.

Surajit Chaudhuri and Moshe Y. Vardi. Optimization of *Real* conjunctive queries. In *Proceedings of the Twelfth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, May 25-28, 1993, Washington, DC, USA*, pages 59–70, 1993. doi: 10.1145/153850.153856. URL `http://doi.acm.org/10.1145/153850.153856`.

Bibliography II

Robert Endre Tarjan and Mihalis Yannakakis. Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM J. Comput.*, 13(3):566–579, 1984. doi: 10.1137/0213035. URL `http://dx.doi.org/10.1137/0213035`.

Mihalis Yannakakis. Algorithms for acyclic database schemes. In *VLDB*, pages 82–94, 1981.