TELECOM ParisTech

Determining Relevance of Accesses at Runtime

Michael Benedikt, Georg Gottlob,
Pierre Senellart

PODS 2011, 14 June 2011

Querying the deep Web

- A large part of deep Web data (phone directories, library catalogs, etc.) is essentially relational
- Access to the deep Web necessary goes through restricted query interfaces, named here access methods
- Typically: for a given form interface to relational data, some input attributes must be bound, other attributes are free
- Given a query (say, conjunctive) over base relations, answering it using restricted interfaces may 1) not be possible 2) require an unbounded number of calls to query interfaces
- Large body of work on the computation of static query plans under access limitations [Rajaraman et al., 1995, Duschka and Levy, 1997, Li, 2003, Nash and Ludäscher, 2004, Calì and Martinenghi, 2008b]: not our concern here

When is an access relevant?

Consider:

- a schema $\mathcal{S}$, with access methods for schema relations
- a query Q over $\mathcal{S}$
- some pre-existing knowledge Conf of the content of relations of $\mathcal{S}$
- an access method over a base relation $R \in \mathcal{S}$, and a binding $\vec{b}$ of the input attributes to constants; the corresponding access is denoted $R(\vec{b}, ? \ldots ?)$ (or $R(\vec{b})$? if there are no output attributes)

We want to know if $R(\vec{b}, ? \ldots ?)$ is relevant to Q in Conf, i.e., if it may bring us knowledge of the truth value of Q.

Motivating example

Schema (input attributes in blue)

```
Employee(EmpId, Title, LastName, FirstName, OffId)
Office(OffId, StreetAddress, State, Phone)
Approval(State, Offering)
Manager(EmpId, EmpId)
```

Query

```
SELECT DISTINCT 1 FROM Employee E, Office O, Approval A
WHERE E.Title='loan officer' AND E.OffId=O.OffId
  AND O.State='Illinois' AND A.State='Illinois' AND A.Offering='30'
```

Is the access "Manager(12345,?)" relevant to the query?

Motivating example

Schema (input attributes in blue)

```
Employee(EmpId, Title, LastName, FirstName, OffId)
Office(OffId, StreetAddress, State, Phone)
Approval(State, Offering)
Manager(EmpId, EmpId)
```

Query

```
SELECT DISTINCT 1 FROM Employee E, Office O, Approval A
WHERE E.Title='loan officer' AND E.OffId=O.OffId
  AND O.State='Illinois' AND A.State='Illinois' AND A.Offering='30'
```

Is the access "Manager(12345,?)" relevant to the query?

Different notions of relevance

The relevance of a =“Manager(12345,?)” depends on several factors:

Initial configuration If we already know of a loan officer in Illinois, a is not relevant. Otherwise, it might be.

Dependence of accesses If it is possible to “guess” employee ids at random (independent accesses), a is not relevant. If all employee ids used must appear as the result of a previous access (dependent accesses), a may be relevant.

Immediate and long-term relevance By itself, a cannot make the query true if it was not true already: it is not immediately relevant. But it may provide employee ids that will be used to build a witness to the query, i.e., it is long-term relevant.

Problem studied

Algorithms for, complexity of determining if an access is relevant to a query in a given configuration:

- independent vs dependent case
- immediate relevance vs long-term relevance
- current access: Boolean (no output attributes) vs non-Boolean
- conjunctive queries (CQs) vs positive queries (PQs)

We focus on combined complexity, but we also present data complexity results.

We relate the notion of access relevance to query containment under access limitations.

Problem studied

Algorithms for, complexity of determining if an access is relevant to a query in a given configuration:

- dependent case
- long-term relevance
- current access: Boolean (no output attributes)
- conjunctive queries (CQs) vs positive queries (PQs)

We focus on combined complexity, but we also present data complexity results.

We relate the notion of access relevance to query containment under access limitations.

Long-term relevance

Query Q, configuration Conf, relation R, tuple $\vec{b}$.

$R(\vec{b})$? is long-term relevant (LTR) for Q in Conf if there exists a path (a valid sequence of subsequent accesses) p such that:

- Conf $+ R(\vec{b}) + p \models Q$
- Conf $+ p \not\models Q$

Outline

Containment under access limitations

Schema $\mathcal{S}$, set of access methods $\mathcal{A}$, configuration Conf.

Definition

Query Q_1 is contained in Q_2 under $\mathcal{A}$ starting from Conf, denoted $Q_1 \sqsubseteq_{\mathcal{A},\mathsf{Conf}} Q_2$ if for every configuration Conf' reachable from Conf,

$$\mathsf{Conf}' \models Q_1 \Rightarrow \mathsf{Conf}' \models Q_2.$$

Notion studied (in a restricted form) in [Calì and Martinenghi, 2008a], shown to be coNEXPTIME for conjunctive queries. No lower bound given.

From containment to relevance and back

Let $\mathcal{Q}$ be one of CQs, PQs.

- There are reductions in both directions between query containment of queries in $\mathcal{Q}$ under access limitations and the complement of LTR of a Boolean access for queries in $\mathcal{Q}$.
- Consequently, upper and lower complexity bounds for containment carry over to LTR.

Outline

Complexity results

Theorem

- *Containment of CQs is coNEXPTIME-complete in combined complexity.*
- *Containment of PQs is co2NEXPTIME-complete in combined complexity.*
- *Containment of PQs is PTIME if the queries are fixed.*

Upper-bound argument

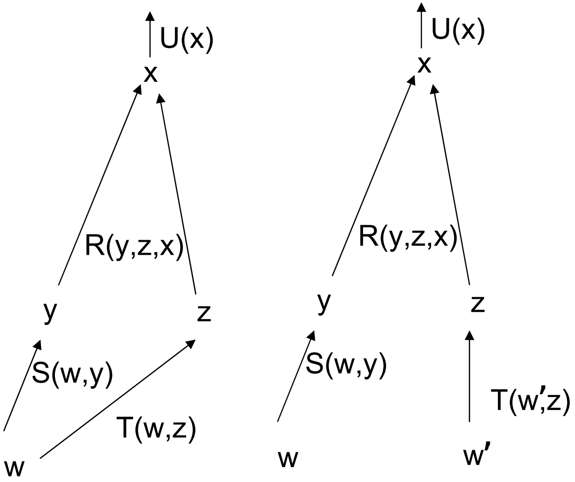
CQ containment under access patterns is a particular case of monadic Datalog containment [Li and Chang, 2001], which yields a 2EXPTIME upper bound [Cosmadakis et al., 1988].

Key arguments for coNEXPTIME (and co2NEXPTIME for PQs):

- A witness instance to non-containment can be made tree-like [Chaudhuri and Vardi, 1997, Calì and Martinenghi, 2008a]: constants produced by an access are used at most once.
- Nodes of a tree-like instance that "have the same type" can be collapsed, reducing the size of the witness.
- For CQs (resp., PQs), nodes have exponentially (resp., doubly exponentially) many possible types.

Tree-likeness

Q= ∃ x U(x) ∧ ....



Lower-bound argument

- Reductions from corridor tiling [Johnson, 1990] under horizontal and vertical constraints
- A tiling will describe a well-formed sequence of accesses, from top-left to bottom-right
- Horizontal and vertical positions are represented through their binary encoding (for PQs, enumerated by an exponential sequence of accesses)
- Queries, together with typing, ensure the path has the required shape, and that constraints are satisfied
- For CQs: $\land$ and $\lor$ encoded with their truth value tables, adding an extra place to relations

Outline

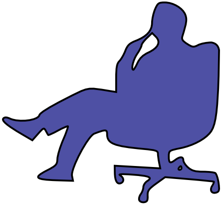
In brief

- Runtime analog of classical problems under access limitations
- Connection between long-term relevance and containment under access limitations
- Combined complexity:

	IR	LTR (Boolean)	Containment
Indep. accesses (CQs)	DP-c	Σ_2^P-c	Π_2^P-c
Indep. accesses (PQs)	DP-c	Σ_2^P-c	Π_2^P-c
Dep. accesses (CQs)	DP-c	NEXPTIME-c	coNEXPTIME-c
Dep. accesses (PQs)	DP-c	2NEXPTIME-c	co2NEXPTIME-c

- Data complexity: everything in PTIME (AC^0 for independent accesses)

Perspectives



- Adding views, integrity constraints, and exactness constraints to the setting (negation)
- Application to runtime optimization of deep Web accesses
- Other notions of relevance:
  - LTR: $\exists$ an instance, $\exists$ a path, such that the query is true after the path and not after the truncation of the path
  - $\exists$ an instance, $\forall$ paths such that the query is true after the path, it is not after the truncation of the path
  - $\forall$ instances, $\exists$ a path, such that the query is true after the path and not after the truncation of the path

Merci.

Webdam

FOX

DIADEM domain-centric intelligent automated data extraction methodology

EPSRC

Pioneering research and skills

Andrea Calì and Davide Martinenghi. Conjunctive query containment under access limitations. In *ER*, 2008a.

Andrea Calì and Davide Martinenghi. Querying data under access limitations. In *ICDE*, 2008b.

Surajit Chaudhuri and Moshe Y. Vardi. On the equivalence of recursive and nonrecursive Datalog programs. *JCSS*, 54(1):61–78, 1997.

Stavros S. Cosmadakis, Haim Gaifman, Paris C. Kanellakis, and Moshe Y. Vardi. Decidable optimization problems for database logic programs. In *STOC*, 1988.

Oliver M. Duschka and Alon Y. Levy. Recursive plans for information gathering. In *IJCAI*, 1997.

David S. Johnson. A catalog of complexity classes. In *Handbook of Theoretical Computer Science*. MIT Press, 1990.

Chen Li. Computing complete answers to queries in the presence of limited access patterns. *VLDB J.*, 12(3):211–227, 2003.

Chen Li and Edward Y. Chang. Answering queries with useful bindings. *ACM Trans. Database Syst.*, 26(3):313–343, 2001.

Alan Nash and Bertram Ludäscher. Processing union of conjunctive queries with negation under limited access patterns. In *EDBT*, 2004.

Anand Rajaraman, Yehoshua Sagiv, and Jeffrey D. Ullman. Answering queries using templates with binding patterns. In *PODS*, 1995.

Framework

We assume given:

- a relational schema $\mathcal{S} = \{S_1 \dots S_n\}$ (each attribute has an abstract domain);
- a set of access methods $\mathcal{A} = \{A_1 \dots A_m\}$ where each A_i is the given of:
  1. one relation S_i of $\mathcal{S}$
  2. a subset of the attributes of S_i that are input attributes
  3. either of the dependent or independent types

Configurations and accesses

- A configuration Conf is an instance of the relational schema.
- Given a configuration Conf, a well-formed access a is the given of:
  - an access method A_k
  - an assignment of input attributes of A_k to constants such that either:
    - A_k is independent
    - or all values of the binding are constants of Conf of the proper domain
- A configuration Conf and a well-formed access a leads (non-deterministically) to a new configuration Conf' with:
  1. $\mathrm{Conf} \subseteq \mathrm{Conf}'$
  2. $\mathrm{Conf}' - \mathrm{Conf}$ only contains tuples of the accessed relation, and all these tuples agree with the binding

Configuration paths

A well-formed path between configurations Conf and Conf′ is a sequence of configurations

$$(\mathsf{Conf} =)\mathsf{Conf}_0 \rightarrow^{a_1} \mathsf{Conf}_1 \rightarrow^{a_2} \dots \quad \mathsf{Conf}_{n-1} \rightarrow^{a_n} \mathsf{Conf}_n (= \mathsf{Conf}')$$

such that for all $i \geqslant 1$, a_i is a well-formed access that leads from Conf_{i-1} to Conf_i. We say Conf′ is reachable from Conf.

The truncation of this path is the path

$$(\mathsf{Conf} =)\mathsf{Conf}_0 \rightarrow^{a_2} \mathsf{Conf}'_2 \rightarrow^{a_3} \dots \quad \mathsf{Conf}_{n-1} \rightarrow^{a_k} \mathsf{Conf}'_k$$

with k maximum such that the path is still well-formed, and Conf'_i contains all facts of Conf_i except those produced by a_1.

Queries

- Only Boolean queries
- Two query languages, subsets of the relational calculus:

  Conjunctive queries (CQs) $\exists, \wedge$

  Positive queries (PQs) $\exists, \wedge, \vee$

- Queries should be consistent with attribute domains
- Constants in the query are assumed to also be part of the configuration
- We note $\mathsf{Conf} \models Q$ when Q is true in Conf

Immediate relevance

Query Q, configuration Conf, access a.

a is immediately relevant (IR) for Q in Conf if there exists a configuration Conf$'$ such that:

- a may lead from Conf to Conf$'$
- Conf $\not\models Q$
- Conf$'$ $\models Q$

Simple example

Example

$Q = R(x, y) \wedge S(y, z)$. $\mathsf{Conf} = \varnothing$. $a = R(?, ?)$. Access method on S.

- a is not IR for Q in Conf.
- a is LTR for Q in Conf.

First observations

- For a fixed arity k, relevance for a query of output arity k reduces to relevance for Boolean queries.
- Determining relevance for Q in Conf requires checking that Conf $\not\models Q$, which is coNP-hard for CQs.

Immediate relevance (independent case)

Proposition

IR for CQs or PQs is DP-complete in combined complexity. If the query is fixed, the problem is in AC^0.

Proof sketch.

Upper bound: the problem is shown to be in NP (by a short-witness argument) as soon as the query is known not to be true.

Lower bound: coding of satisfiability/unsatisfiability pair as a single query.

Data complexity: the algorithm can be implemented as a first-order formula.

Immediate relevance (independent case)

Proposition

IR for CQs or PQs is DP-complete in combined complexity. If the query is fixed, the problem is in AC^0.

Proof sketch.

Upper bound: the problem is shown to be in NP (by a short-witness argument) as soon as the query is known not to be true.

Lower bound: coding of satisfiability/unsatisfiability pair as a single query.

Data complexity: the algorithm can be implemented as a first-order formula.

Long-term relevance (independent case)

Proposition

In the absence of dependent accesses, the combined complexity of LTR for CQs or PQs is Σ_2^P*-complete. If the query is fixed, the problem is in* AC^0.

Proof sketch.

The upper bound is straightforward. The lower bound is a consequence of a the hardness of determining whether a tuple is critical for a query in a relational database [?]. □

Long-term relevance (independent case)

Proposition

In the absence of dependent accesses, the combined complexity of LTR for CQs or PQs is Σ_2^P*-complete. If the query is fixed, the problem is in* AC^0.

Proof sketch.

The upper bound is straightforward. The lower bound is a consequence of a the hardness of determining whether a tuple is critical for a query in a relational database [**?**]. □

From containment to relevance

Let $\mathcal{Q}$ be one of CQs, PQs.

Proposition

There is a polynomial-time many-one reduction from query containment of queries in $\mathcal{Q}$ under access limitations to the complement of LTR of dependent accesses for queries in $\mathcal{Q}$.

From relevance to containment

Proposition

There is a reduction from LTR of dependent Boolean accesses to the complement of query containment, which is:

- *a polynomial-time many-one reduction for PQs;*
- *a nondeterministic polynomial-time Turing reduction for CQs.*

The weaker form of reduction comes from the need for disjunction. Enough to show matching complexity results for containment and LTR (in the Boolean case).