Communication in Federated Query Evaluation

Tamara Cucumides
University of Antwerp

Juan Reutter D.
Universidad Católica de Chile
Instituto Milenio Fundamentos de los Datos
Web Query Federation

Query → Federation Engine

Connection via http or similar

Data Repository

Data Repository

Data Repository

Federation Engine does not control the data!
Web Query Federation

Best activities for today in Santiago

Best activities in Santiago

Weather in Santiago

Tweets under #Santiago, #Fun

Tripadvisor

Weather Data

Twitter
Web Query Federation

- SPARQL has a standard to do this

```sparql
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name
FROM <http://example.org/myfoaf.rdf>
WHERE
{
  SERVICE <http://people.example.org/sparql> {
    ?person foaf:name ?name .
  }
}
```
Web Query Federation

- SPARQL has a standard to do this
- How to evaluate those queries? Good body of work researching this issues
- Most of it: heuristically search for best query, or construct better infrastructure
Web Query Federation

- SPARQL has a standard to do this
- How to evaluate those queries? Good body of work researching this issues
- Most of it: heuristically search for best query, or construct better infrastructure

What is the communication cost of these queries?
Can we find provable optimal algorithms?
Question 1: What is the minimum amount of bits needed to answer federated queries?

Question 2: Can we build algorithms communicating this amount?
Our setting

\[ Q(x, y, z) \leftarrow R(x, y) \land S(x, z) \land T(y, z) \]
Question 1: What is the minimum amount of bits needed to answer federated queries?

Question 2: Can we build algorithms communicating this amount?

Yes! Linear amount, send all the data!
Definitely not a good algorithm

But optimal in terms of size of relations
Question 1: What is the minimum amount of bits needed to answer federated queries?

Question 2: Can we build algorithms communicating this amount?

• Need more fine-grained measures. Distinguish hard and easy instances.
• Focus on adaptive complexity!
Adaptive complexity - set intersection

- Sort-merge intersection
Adaptive complexity - set intersection

- Sort-merge intersection
Adaptive complexity - set intersection

- Sort-merge intersection
Adaptive complexity - set intersection

- Sort-merge intersection
Adaptive complexity - set intersection

- Sort-merge intersection
Adaptive complexity - set intersection

- Sort-merge intersection
Adaptive complexity - set intersection

- Sort-merge intersection
Adaptive complexity - set intersection

- Sort-merge intersection
Adaptive complexity - set intersection

- Sort-merge intersection
- Linear complexity, BUT
- Time = number of changes
Adaptive complexity - set intersection

- Sort-merge intersection can intersect two sets communicating only numbers bounded by the number of changes.
- Better than sending all the data!
- Also: almost optimal in terms of communication complexity, for an instance with $n$ elements and $c$ changes, the communication complexity is

$$\Omega(\log \left( \binom{n}{c/2} \right)) \sim c \log n$$
Adaptive complexity - relational data

- Notion of certificate from Ngo et al. 2014

\[ R[0,3] : \] lowest value on first component, third lowest in second

Certificate: set of comparisons

\[
\begin{align*}
R[\bar{a}] &< S[\bar{b}] \\
R[\bar{a}] &= S[\bar{b}]
\end{align*}
\]

Such that the value of a query is fully determined by these comparisons
Certificates (Ngo et al. 2014)

\[ Q(x, y, z) \leftarrow R(x, y) \land S(x, z) \land T(y, z) \]

Certificate for this query:

\[
\begin{align*}
R[0] &= R[0,0] \\
S[0] &= S[0,0] \\
T[0] &= T[0,0] \\
S[0, -1] &= S[0,0] \\
S[-1] &< R[1] \\
T[-1] &< R[0,1]
\end{align*}
\]
Certificates (Ngo et al. 2014)

\[ Q(x, y, z) \leftarrow R(x, y) \land S(x, z) \land T(y, z) \]

Certificate for this query:

\[
\begin{align*}
R[0] &= R[0,0] \\
S[0] &= S[0,0] \\
T[0] &= T[0,0]
\end{align*}
\]

\[
\begin{align*}
R[0] &= S[0] \\
S[0] &= T[0]
\end{align*}
\]

Accounts for \((1,1,1)\) as an answer.
Certificates (Ngo et al. 2014)

\[ Q(x, y, z) \leftarrow R(x, y) \land S(x, z) \land T(y, z) \]

Certificate for this query:

\[ S[0, -1] = S[0,0] \]

Invalidates any other answer of form \((1,1,b)\), \(b > 1\)
Certificates (Ngo et al. 2014)

\[ Q(x, y, z) \leftarrow R(x, y) \land S(x, z) \land T(y, z) \]

Certificate for this query:

\[ S[-1] < R[1] \]

Invalidates any other answer of form \((b, , )\), \(b > 1\)
Certificates (Ngo et al. 2014)

\[ Q(x, y, z) \leftarrow R(x, y) \land S(x, z) \land T(y, z) \]

Certificate for this query:

Invalidates any other answer of form \((1, b, )\), \(b > 1\)

\[ T[-1] < R[0,1] \]
Adaptive complexity - relational data

- What is the communication cost for federated queries in terms of certificates?

Proposition.
If $Q$ is acyclic, and $I$ has $n$ elements and certificate size $c \leq \frac{n}{2}$, then

Any deterministic algorithm must communicate $\Omega(c \log n)$ bits.
Adaptive complexity - relational data

• What is the communication cost for federated queries in terms of certificates?

Proposition.
For every $n, c > 1$, there is
- query $Q$,
- instance $I$ with $n$ elements and certificate size at most $c$,
Any deterministic algorithm must communicate $\Omega(c \log n)$ bits.
Adaptive algorithm - relational data

- We also have a matching upper bound!
- Builds from Minesweeper (Ngo et al. 2014)
- Every server communicates a number of tuples in $O(c + |R \bowtie Q(I)|)$
\[ Q(x, y, z) \leftarrow R(x, y) \land S(x, z) \land T(y, z) \]
$Q(x, y, z) \leftarrow R(x, y) \land S(x, z) \land T(y, z)$

$\begin{array}{ccc}
R & S & T \\
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
\end{array}$

$\begin{array}{ccc}
\text{yes} & \text{yes} & \text{yes} \\
1 & 1 & 1 \\
\end{array}$

(1,1,1)?
\[ Q(x, y, z) \leftarrow R(x, y) \land S(x, z) \land T(y, z) \]
\[ Q(x, y, z) \leftarrow R(x, y) \land S(x, z) \land T(y, z) \]

\[ \langle 1, (2, 5) \rangle \quad \langle 1, (2, \infty) \rangle \]
\[ Q(x, y, z) \leftarrow R(x, y) \land S(x, z) \land T(y, z) \]

\[ \langle 1, (2, 5) \rangle \quad \langle 1, (2, \infty) \rangle \]

\[ (1,2,1) \]

\[ (1,2) \in R? \]

\[ (2,1) \in T? \]
Wrapping Up

*Adaptive complexity* seems like a nice framework to study communication in web query federation.

So far we have studied upper and lower bounds for join queries.
More to do!

Does certificates make sense in real life?

How does our algorithm fares in federated query benchmarks?

Randomised communication algorithms? More queries?
Communication in Federated Query Evaluation

Tamara Cucumides

Juan Reutter D.

University of Antwerp

Universidad Católica de Chile
Instituto Milenio Fundamentos de los Datos