CC71X : Linked Data

Juan Enrique Muñoz Zolotoochin

Queries and Computation on the Web

Serge Abiteboul and Victor

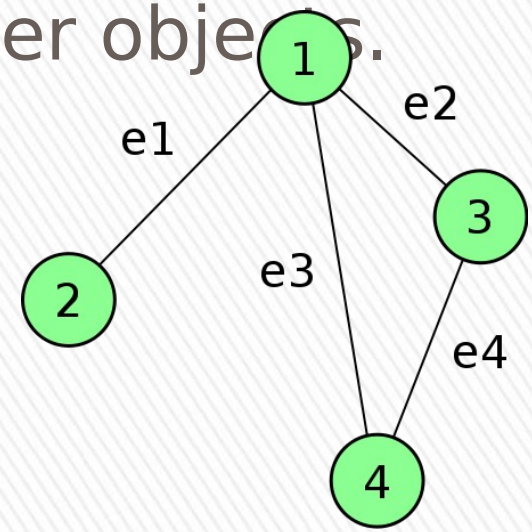
Introduction

- **The web**: tremendous source of information.
- Can be viewed as a large, loosely structured database .
- So can we query it like a database?. First we need:
  - A model of the web.
  - A suitable query language.
- Which queries are actually computable?

The Model

- **Infinite** web : exhaustive exploration is unfeasible.
- It is an infinite set of **objects** which have a **value** and **labeled references** to other objects.

- **INTUITION:**
  - Objects are web pages.
  - References are links.



» **Formaly** – An infinite database over the fixed relational schema:

{ Obj(oid), Ref(source, label, destination), Val(oid, value)}

» Which satisfies:

- $Obj = \pi_{source}(Ref) \cup \pi_{oid}(Val)$
- $\pi_{dest}(Ref) \subseteq Obj$
- $\forall o \in Obj\ \sigma_{source=o}(Ref)$ is finite.
- At most one value per object.

The model

›

- We want to compute queries over web instances.
- Queries are **mappings** from a web instances I to a subset of I(Obj).
- **Generic** query: *q* is generic if for each web instance *I* and one-to-one mapping ρ : **q(ρ(I))=ρ(q(I))**
-  INTUITON : The result only depends on the information of I and does not depend on the encoding.

Computability

- **Web Machine** : Like a Turing Machine but with…
  - A right-infinite input tape.
  - A two-way-infinite work tape.
  - A right-infinite output tape.
- **Initial State** : input tape contains an encoding of the Web Instance.
- Output head can only move forward (nothing is deleted).

Computability

Computability

- Computability requires an initial set of known objects.

- **Computable**: There exists a WM which on input *enc(I)* halts and produces *enc(q(I)).*
- **Eventually Computable**: Exists a WM for which
  - The content of the output tape is always a prefix of ***enc(q(I))***
  - Each ***o*** in ***q(I)*** occurs at some point in the computation.

Computability

- Computable
  - Find the objects **o'** such that there is a path of length at most k from **o** to **o'**.
- Eventually Computable, possibly infinite answers
  - Find the objects reachable from **o**.
- Eventually Computable, finite answers
  - Find the objects on the shortest cycle containing **o**.
- Not eventually computable
  - Find all objects which are not referenced by any object.

Browse & Search

- In practice we use two modes of computation: **Browsing** and **Searching**.
- **Browse Machine**, like a **Web Machin**e but with an *expand* state and a browse tape instead of input tape.
- Initially, the browse tape contains the encoding of an initial object **o**.
- **Expands** replaces the browsing tape with the encoding of all nodes referenced in it.

THEOREM

"Every generic and computable Web Query is browser computable."

Browse & Search ›

Browse & Search

- **Browse/Search Machine**, a **Browse Machine** with
- A search-answer tape.
- A search-condition tape.

- Conditions involve a finite set of (in)equalities involving an attribute and a constant.

$\sigma_{value=uchile}(Val)$:

"Search for all tuples *Val(o, 'uchile')*"

THEOREM

“A generic Web query is eventually computable iff it is eventually computable by a browse/search machine. ”

“A generic Web query is computable iff it is computable by a browse/search machine”

Browse & Search

Query Languages

- We consider **FO** (First-order logic), **FO+** (without negation), **Datalog** and **Datalog¬** (with negation).

- Are the queries on each language (eventually) computable?
- Which fragment of each language can be implemented by browsers?

THEOREM

"All FO+ and Datalog queries are eventually computable."

Query Languages

- Source-ranged-restricted variables
  - If **R(u)** occurs in the body of the rule, **R** is some idb predicate and **x** is one of the variables of **u**, then **x** is s*ource-range-restricted*.
  - If **x** is the source constant or x is source-range-restricted and **Ref(x,y,z)** appears in the body of the rule, then **y** and **z** are *source-range-restricted*.
  - If **x** is the source constant or **x** is source-range-restricted and **Val(x,y)** occurs in the body, **y** is *source-range-restricted.*

Query Languages ›

» **Source safe:**

answer(source) ←

answer(‘t) ← answer(t), Ref(t,x,t’)

» **Not source safe:**

answer(source) ←

answer(‘t) ← answer(t’), Ref(t,x,t’)

Query Languages ›

THEOREM

"All ss-FO queries are computable by a browser machine."

"All ss-Datalog queries are eventually computable by a browser machine."

- Datalog ¬, problems arise with negation:
  - P ← ¬ P
  - Single(x) ← Man(x), ¬ Married(x)
  - Married(x) ← Man(x), ¬ Single(x)

- Has different semantics
  - Well-founded semantics
  - Stratified semantics
  - Inflationary semantics

Query Languages

THEOREM

“Every query in ss-Datalog¬ with inflationary semantics is eventually computable by a browser machine. ”

Query Languages

Conclusion :

ss-Datalog¬ with inflationary semantics emerges as a particularly appealing language in the context of the web.