WEBPIE: A WEB-SCALE PARALLEL INFERENCE ENGINE

Jacopo Urbani, Spyros Kotoulas, Jason Maassen, Niels Drost, Frank Seinstra, Frank van Harmelen, Henri Bal

Department of Computer Science
Vrije Universiteit Amsterdam
The Semantic Web

- The Semantic Web is an extension of the current Web where the semantics is defined.

- Basically the idea is to move from

  **Web of Documents**
  (Traditional Web)

  to

  **Web of data**
  (Semantic Web)
The Semantic Web

- In Semantic Web the data is written in RDF
The Semantic Web

- In Semantic Web the data is written in RDF

  `<http://www.vu.nl> <rdf:type> <http://university.com>`
The Semantic Web

- In Semantic Web the data is written in RDF
  
  `<http://www.vu.nl> <rdf:type> <http://university.com>`

- Machines can apply rules and derive new statements. We call it reasoning
In Semantic Web the data is written in RDF

```
<http://www.vu.nl>  <rdf:type>  <http://university.com>
```

Machines can apply rules and derive new statements. We call it reasoning.

**Input:**

```
<Jacopo>  <type>  <Student>
<Student>  <subclass>  <Person>
```

**Rule to apply:**

```
if a type B and B subclass C then a type C
```

**Output:**

```
<Jacopo>  <type>  <Person>
```
The Semantic Web

Advantages:

- able to combine data from different documents
The Semantic Web

Advantages:
- able to combine data from different documents
- answer to complex queries
  - “list of the first ten peaks in Europe”
  - “find all publications on grid computing since 1995”
The Semantic Web

Advantages:

- able to combine data from different documents
- answer to complex queries
  - “list of the first ten peaks in Europe”
  - “find all publications on grid computing since 1995”
- return also derived information
  - ask for persons -> the system returns also the students
Web scale reasoning

- Size of the Semantic Web
  - March 2009: 4.4 Billions
  - Sept. 2009: 7.7 Billions
  - Jan. 2010: 13.1 Billions
  - Now: ?!?!

- Input size → need for parallelization
- Explosive growth → need for scalability
Web scale reasoning

- Size of the Semantic Web
  - March 2009: 4.4 Billion
  - Sept. 2009: 7.7 Billion
  - Jan. 2010: 13.1 Billion
  - Now: ?!!

- Input size → need for parallelization
- Explosive growth → need for scalability

Solution: WebPIE!!

Monday 10 May 2010
WebPIE – what is it?

WebPIE is a MapReduce distributed reasoner that works on a Web scale

Features:
- High performance
- High scalability
WebPIE – what is it?

WebPIE is a MapReduce distributed reasoner that works on a Web scale

Features:
- High performance
- High scalability

Reason on the entire Semantic Web

60 times faster!!!
WebPIE – implementation

- Straightforward MapReduce is slower than sequential program
  - Load balancing
  - Duplicate derivations
  - etc.

- WebPIE introduces novel techniques that, combined, solve all the issues
  (for details, see the papers)
WebPIE – implementation

- Written in Java, uses Hadoop (0.20.2)
- Run on cluster or on the Amazon cloud
- Tests performed at the DAS3 cluster
  http://www.cs.vu.nl/das3
- Code, tutorial, etc. available at
  http://www.cs.vu.nl/webpie
Performance

![Graph showing performance comparison between different systems. The x-axis represents input size in billions of statements, and the y-axis represents throughput in KTriples/sec. The graph shows data for BigOWLIM, Oracle 11g, DAML DB, and BigData.]

Monday 10 May 2010
Performance

We are here!!

Monday 10 May 2010
## Performance

- Tested on different datasets
  - different input size
  - different input complexity
- In all cases the performance is better than best technique

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Input size</th>
<th>Output size</th>
<th>Exec. time</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUBM</td>
<td>1 Billion</td>
<td>0.5 Billion</td>
<td>0.6 hours</td>
</tr>
<tr>
<td>Uniprot</td>
<td>1.5 Billion</td>
<td>2.0 Billion</td>
<td>6.1 hours</td>
</tr>
<tr>
<td>LDSR</td>
<td>0.9 Billion</td>
<td>0.9 Billion</td>
<td>3.5 hours</td>
</tr>
</tbody>
</table>
Scalability

Scalability on the input size using 64 nodes

![Graph showing the relationship between input size and execution time](image-url)
Scalability

- Scalability on the number of nodes (input: 10 billions statements)
Conclusions

- Vastly outperforms current state of the art
  - one order of magnitude input size
  - Throughput between 5 and 60 times higher

“WebPIE makes reasoning over the entire Semantic Web possible”
Demo