Data-oriented Content Query System: Searching for Data into Text on the Web

Mianwei Zhou, Tao Cheng, Kevin Chen-Chuan Chang
WSDM 2010, New York, USA
Many Web Applications Try to Exploit the “Content” of Web Pages.

In most cases, what we really want are not pages, but the information units inside.

*Web Info Extraction*  *Typed Entity Search*  *Web-based Q/A*
Content-Exploiting Application 1: Web Information Extraction

Web Information Extraction (WIE)

Pattern:
“X is CEO of Y”

<table>
<thead>
<tr>
<th>Company</th>
<th>CEO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>Eric Schmidt</td>
</tr>
</tbody>
</table>

Limitation
- Focus on simple patterns.
- Lack of interactivity.
Content-Exploiting Application 2: Web-based Question Answering

Web-based Question Answering (WQA)

Who is CEO of Dell?

Limitation
• Only rely on top-k pages to retrieve the answer.

Michael Dell
Typed-Entity Search (TES)

Limitation
• Limited Number of Data Type
• Lack of Flexibility

But ... Where is CEO?
General System for Querying Text “Content”, Much Like How DBMS Supports Data Application

Requirements
1. Extensible Data Types
2. Flexible Contextual Patterns
3. Customizable Scoring
Input: CQL (Content Query Language)

Data-oriented Content Query System

Output: Table
Demo

Online Demo

http://wisdm.cs.uiuc.edu/demos/docqs
Why Relational Model: Occurrence->Tuple

What we need

Person
Organization
Location
Number

Relational Model

Number

Person
Organization
Location
Why Relational Model: Content Query->Relational Operation

What we need

Find the population of China

Relational Model

FROM #number
WHERE pattern(…)
GROUP BY #number
ORDER BY conf()

China has a population of 1.3 billion
China with its population of 1.3 billion people
China is established in 1949
Shanghai is the largest city with 15 million inhabitants in China

1. 1.3 billion
2. 15 million
…
Why Relational Model Extensible Data Type -> View

What we need

- Number
  - Population
    - Price
    - Phone
  - Capital
    - Headquarter
  - Person
    - CEO
    - Professor
    - President

Relational Model

Table

<table>
<thead>
<tr>
<th>Number</th>
<th>Population</th>
<th>Price</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

View

- population
- price
- phone
Main Technique: Highly Efficient and Scalable Index Structure

Data Type Definition

**INPUT**

SELECT ...
FROM ...
WHERE ...

**OUTPUT**

```
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
</table>
```

**Experimental Result**

- *Speed improvement*: 6-10 times
- *Space overhead*: Around 2 times original corpus size.

Index Layer

Repository

Data Type: Definition

- Contextual Index
- Join Index

Query Optimization

Graph Coverage Problem
Data-oriented Content Query System: Supporting Ad-hoc, Interactive “Content” Search.

Web Info Extraction
Typed Entity Search
Web-based Q/A

Data-oriented Content Query System
Interactive
Ad-hoc
Q & A
Thank You!