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.

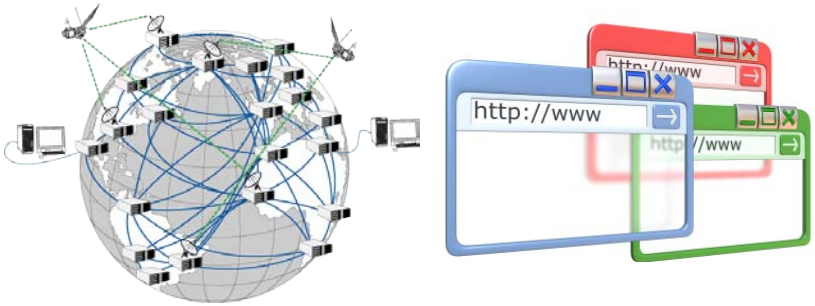
Web Info Extraction

Typed Entity Search

Web-based Q/A

○○○

? ?



Content-Exploiting Application 1:
Web Information Extraction

Web Information Extraction (WIE)

(Marius 2006, Cafarella 2005, Etzioni 2004)

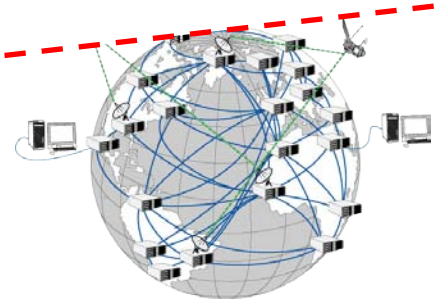
Company	CEO
Google	Eric Schmidt

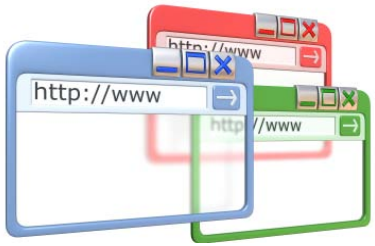
Pattern:

"X is CEO of Y"

Limitation

- Focus on simple patterns.
- Lack of interactivity.





Content-Exploiting Application 2:
Web-based Question Answering

Web-based Question Answering (WQA)

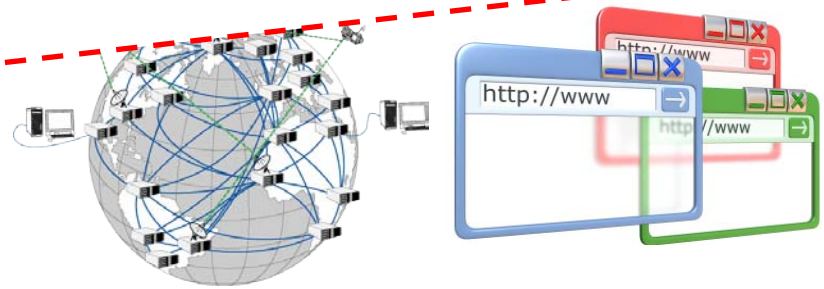
(Wu 2007, Lin 2003, Brill 2002)

Who is CEO of Dell?

Michael D...

Limitation

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



Content-Exploiting Application 3:
Typed-Entity Search

Typed-Entity Search (TES)

(Cheng 2007, Cafarella 2007, Chakrabarti 2006*)*

Entity Search

Ranked Entity List

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

Web Info Extraction

Typed Entity Search

Web-based QA

○○○

Data-oriented Content Query System

Requirements

1. Extensible Data Types
2. Flexible Contextual Patterns
3. Customizable Scoring

System Framework

Input: CQL Output: Table

Common Users

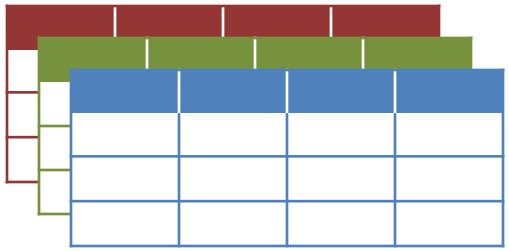
Expert Users

Entity Search

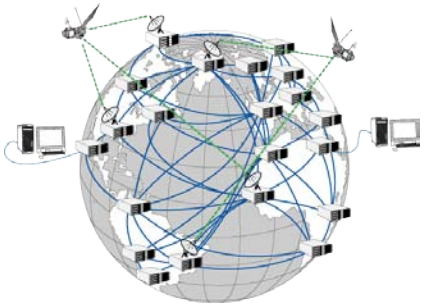
Web QA

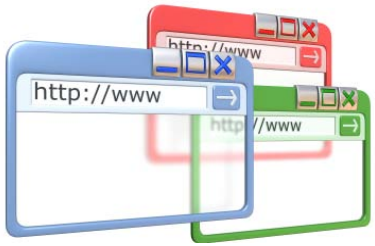
Input: CQL (Content Query Language)

Output



Data-oriented Content Query System





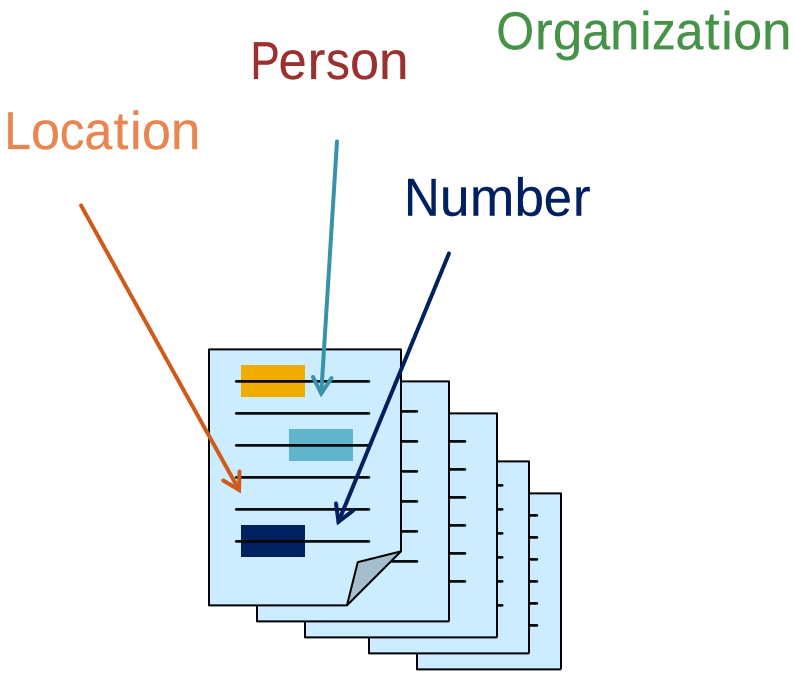
Demo

Online Demo

http://wisdm.cs.uiuc.edu/demos/docqs

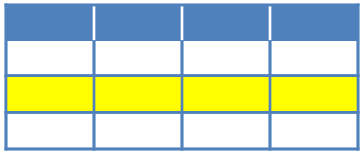
Why Relational Model: Occurrence->Tuple

What we need



Relational Model

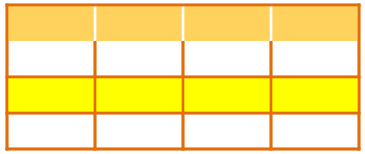
Number



Person

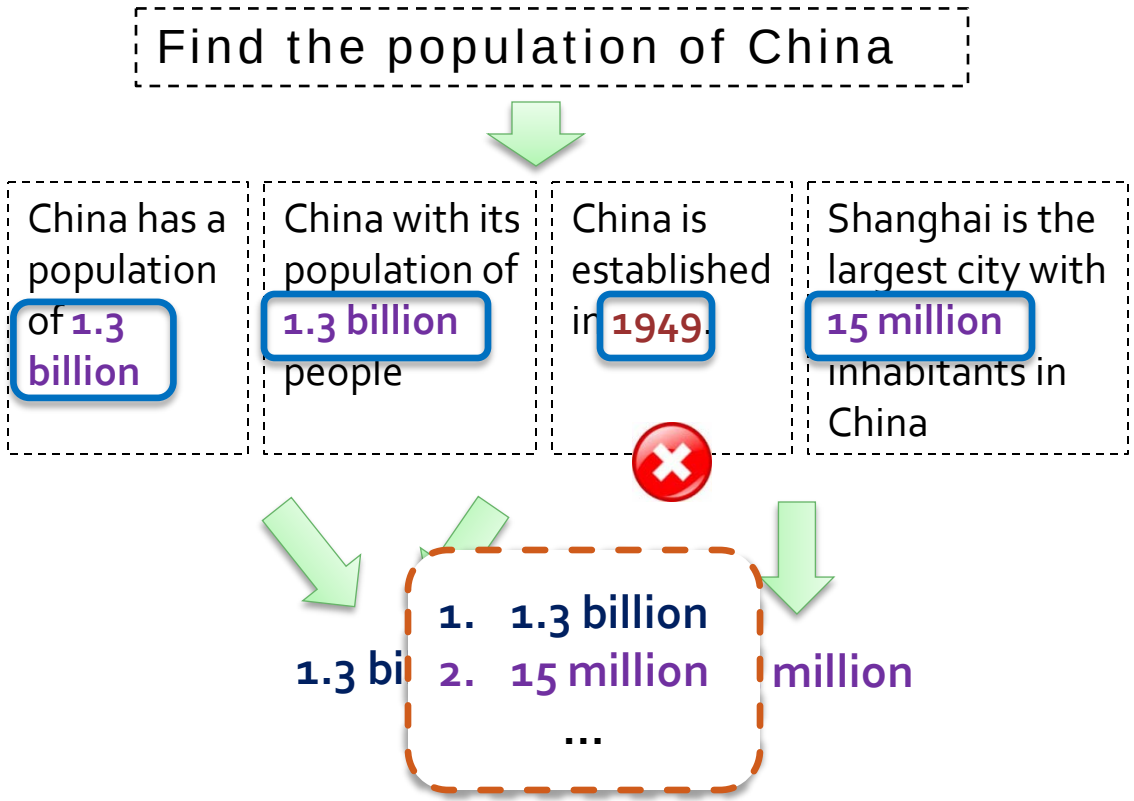
Organization

Location



Why Relational Model: Content Query->Relational Operation

What we need



Relational Model

FROM #number

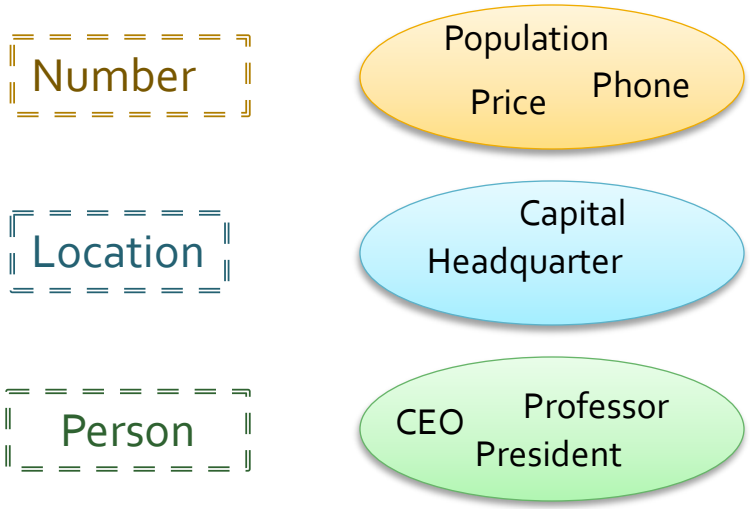
WHERE pattern(...)

GROUP BY #number

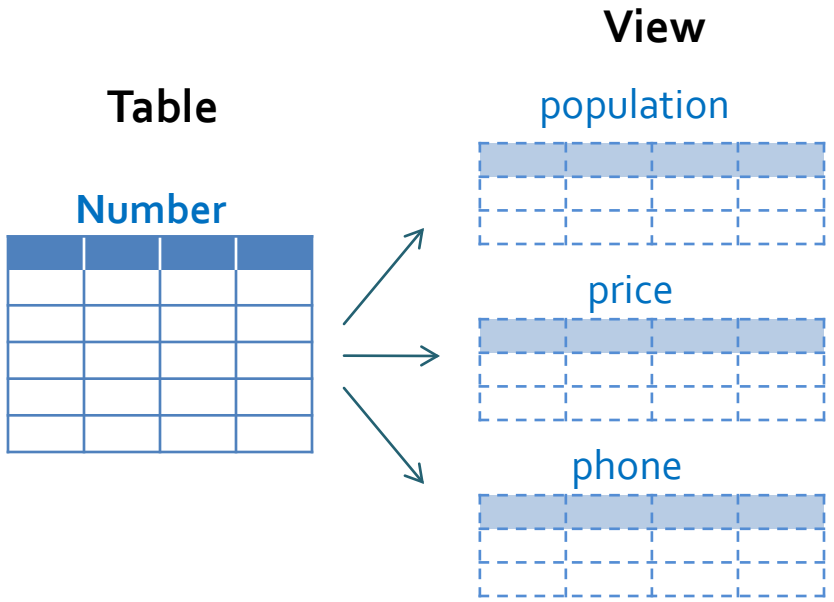
ORDER BY conf()

Why Relational Model Extensible Data Type->View

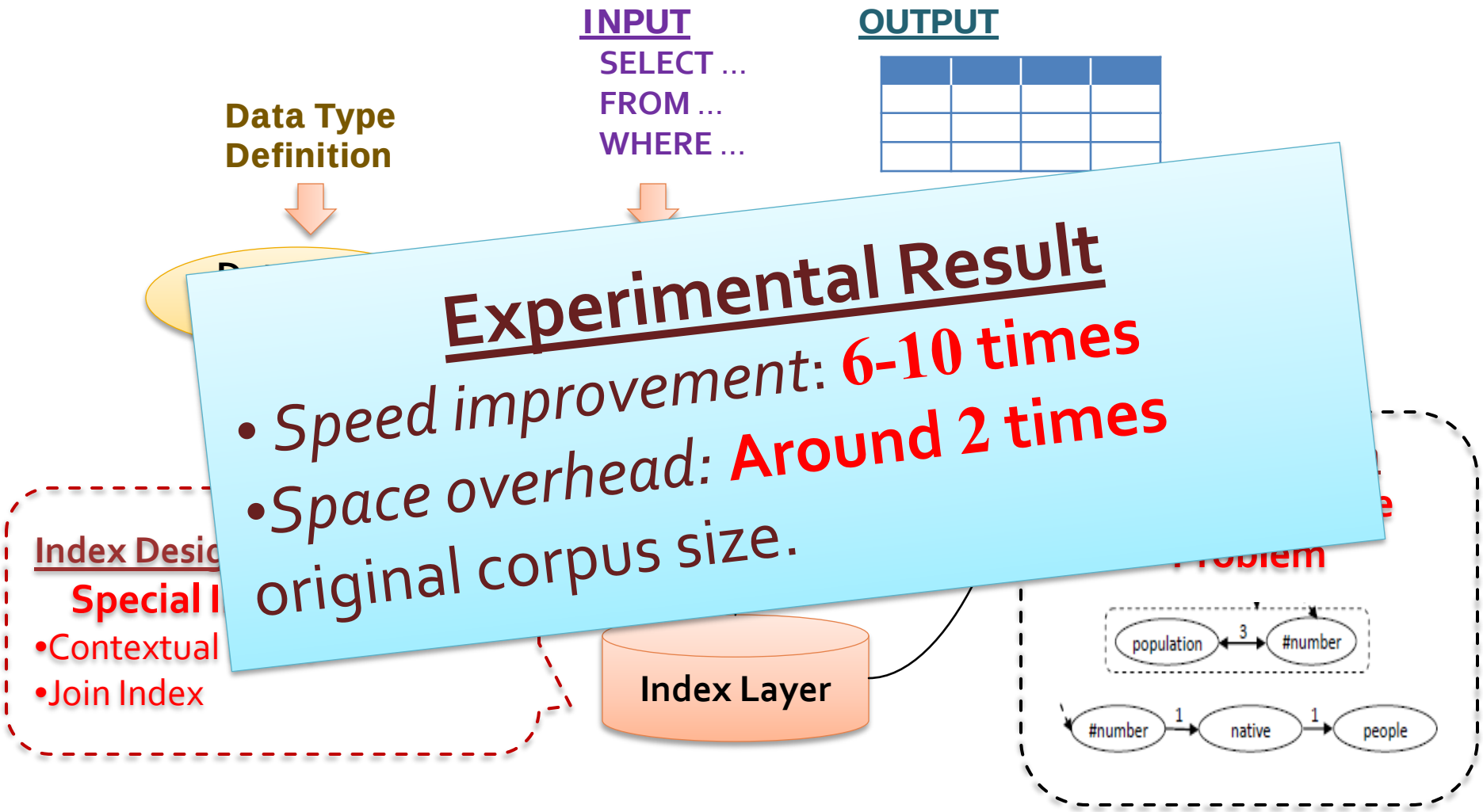
What we need



Relational Model



Main Technique: Highly Efficient and Scalable Index Structure



Data-oriented Content Query System: Supporting Ad-hoc, Interactive "Content" Search.

Web Info Extraction

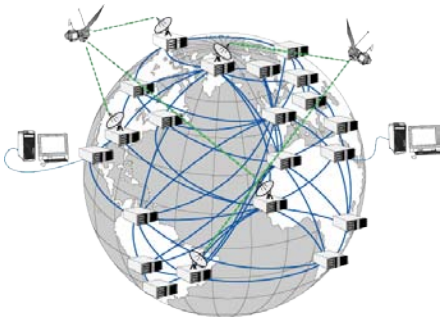
Typed Entity Search

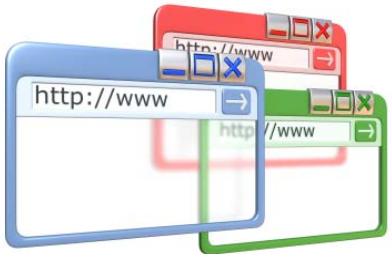
Web-based Q/A

○○○

Data-oriented Content Query System

Interactive Ad-hoc





Q & A

Thank You!