Building Data Integration Systems via Mass Collaboration

AnHai Doan
Dept. of Computer Science
Univ. of Illinois, Urbana-Champaign

Joint work with Robert McCann, Vanitha Varadarajan, & Alexander Kramnik

WebDB 2003
Architecture of Data Integration System

Find books written by Isaac Asimov & priced under $15

source schema 1

source schema 2

source schema 3

mediated schema

wrapper

amazon.com

bn.com

powell.com
Current State of Affairs

- **Vibrant** research & industrial landscape
- **Research**
  - dated back to the 70-80s, accelerated in recent years
  - focused on
    - conceptual & algorithmic aspects
    - building specialized systems
- **Industry**
  - more than 50 startups in 2001

Despite much R&D activities, however …
... DI Systems still Incur Very High Cost of Ownership!

- Most systems are still deployed manually by system admins
  - construct mediated- & source schemas
  - build wrappers
  - find semantic mappings between schemas
  - monitor & adjust to changes at sources

- Manual deployment is extremely labor-intensive
  - now a key bottleneck to widespread deployment

- Emerging technologies (XML, Web services, Semantic Web) will further fuel DI applications & exacerbate the problem

Reducing cost of ownership for DI apps is now crucial!
The MOBS Project

- **MOBS** = Mass Collaboration to Build Systems
- Key idea: spread burden thinly over a mass of users

- treat a DI system as **having a finite set of parameters**
- system admins construct and deploy a system “shell”
- users **help system “converge”** to correct parameter values
## Example: Schema Matching

**PARAMETER:** (author = writer)  
**VALUE:** ?

<table>
<thead>
<tr>
<th>writer</th>
<th>amount</th>
<th>publisher</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Stu</td>
<td>$14.99</td>
<td>Newsday</td>
</tr>
<tr>
<td>......</td>
<td>......</td>
<td>......</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>author</th>
<th>cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upton Sinclair</td>
<td>$6.99</td>
</tr>
<tr>
<td>......</td>
<td>......</td>
</tr>
</tbody>
</table>

**PARAMETER:** (author = writer)  
**VALUE:** yes

<table>
<thead>
<tr>
<th>writer</th>
<th>amount</th>
<th>publisher</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Stu</td>
<td>$14.99</td>
<td>Newsday</td>
</tr>
<tr>
<td>......</td>
<td>......</td>
<td>......</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>author</th>
<th>cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upton Sinclair</td>
<td>$6.99</td>
</tr>
<tr>
<td>......</td>
<td>......</td>
</tr>
</tbody>
</table>
Comparison to Database Tuning

● Database tuning
  – set values of physical-design knobs (e.g., buffer size)
  – using feedback from query execution
    – time, resources consumed, etc.
  – to further improve query execution performance

● Mass collaboration for DI systems
  – set values of logical-design knobs (e.g., "a = b")
  – using feedback from users
  – to improve system correctness and further expand system
Potential High Impact

● If succeeds
  – dramatically reduce cost & time
  – launch numerous DI systems on Web & enterprises
    – everyday domains: books, movies, cars, travel, etc.
    – "niche" domains: e.g., fire fighting
    – scientific domains: e.g., bioinformatics
    – within/across enterprises
  – applicable to other data management tasks
    – building P2P systems, info extraction from text, Semantic Web, ...

● Our current work
  – start by exploring a simple setting:
    mass collaboration to find 1-1 semantic mappings
  – use the setting to understand key challenges
  – develop, deploy, & evaluate general solutions
1. Build a Partial Correct System

Mediated Schema

title author year price category

Schema $S_1$  
$a_1$  $a_2$  $a_3$  $a_4$  $a_5$

Schema $S_2$  
$b_1$  $b_2$

Schema $S_3$  
$c_1$  $c_2$  $c_3$

Schema $S_4$  
$d_1$  $d_2$  $d_3$  $d_4$  $d_5$  $d_6$

SEARCH FOR BOOKS

Enter Title: Mrs. Dalloway

Enter Author: 

Submit Query
2. Solicit User Feedback

- Login: [Input]
- Submit

SEARCH FOR BOOKS

Enter Title: [Input]
Enter Author: [Input]
Submit Query

SEARCH RESULTS
Mrs. Dalloway Virginia Woolf 1990 7.20 Classic Buy
Mrs. Dalloway Virginia Woolf 1992 16.00 BestSeller Buy
Mrs. Dalloway Virginia Woolf 2002 18.00 BestSeller Buy
Mrs. Dalloway Virginia Woolf 1990 12.00 Classic Buy

Back to QueryPage

Is the Highlighted Column a Year?

<table>
<thead>
<tr>
<th>Price</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.45</td>
<td>Classic</td>
</tr>
<tr>
<td>6.50</td>
<td>Children's</td>
</tr>
<tr>
<td>10.00</td>
<td>Classic</td>
</tr>
<tr>
<td>10.00</td>
<td>Classic</td>
</tr>
</tbody>
</table>

Yes No

Context information
Detect & Remove Bad Users

- Insert questions whose answers we already know
- Evaluate user trustworthiness on those questions
- Ignore users with low trustworthy value
3. Combine User Feedback
Empirical Evaluation

- **Simulation**
  - 5000 users, 10 sources, 10 mediated-schema attributes
  - system admin must do work that amounts to **1000 questions**
  - with mass collaboration: each user answers on average **14 questions**
    - burden can be spread thinly over a mass of users

- **Real data + real user experiments in book domain**
  - varying settings with 8 - 11 people
  - some people intentionally provided wrong answers
  - system quickly converge to correct values
    - real users can handle cognitive load of questions in this domain and quickly answer them
Key Challenges

- How to entice users to answer questions?
  - build a partial system, ask user to "pay" when using it
  - channel "payments" from other systems, provide incentives

- What types of questions to ask?
  - cognitively simple, can be answered quickly

- Can DI tasks be broken down into series of such questions?
  - it appears that many tasks can

- How to detect malicious/ignorant users
  - evaluate on questions with known answers

- How to combine user answers
  - use learning/statistical techniques
Related Work

- **Mass collaboration**
  - product review websites [amazon.com, epinions.com, etc.]
  - proposed to build knowledge bases [Richardson & Domingos 03], tech support websites [Ramakrishnan, quiq.com], user trust on Semantic Web [Richardson et. al. 03]
  - first to propose mass collab. for building systems

- **Building data integration systems**
  - many works on reducing cost of specific tasks
  - few on reducing cost of whole process [Rosenthal et. al. 01]

- **Autonomic systems**
  - mass collab. gives DI systems autonomic properties

- **Database tuning, information extraction, Semantic Web**
Conclusion

- Manual deployment is extremely labor-intensive
  - a key bottleneck to widespread deployment of DI systems

- We proposed the MOBS solution
  - lift the enormous burden of system deployment from admins
  - spread it thinly over a mass of users
  - developed & evaluated solutions for a simple DI setting
  - exploring key challenges and proposed solutions

- Future work
  - explore complex schema matching, other DI tasks
  - develop, deploy, and evaluate general solutions
  - examine applicability to other data management tasks

See paper and "anhai on google" for more info.