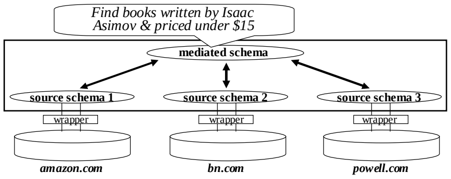
Learning From Multiple Users to Improve Accuracy of Data Integration Tasks

Robert McCann, Alexander Kramnik, Warren Shen, Vanitha Varadarajan, Olu Sobulo, AnHai Doan

DAIS The Database and Information Systems Laboratory at The University of Illinois at Urbana-Champaign Large Scale Information Management

University of Illinois @ Urbana

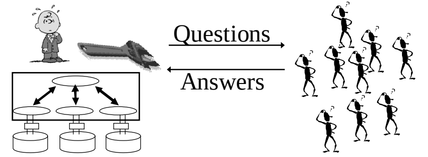
High Cost of Data Integration Systems



- Builder must execute multiple tasks
  - source discovery, wrapper construction, schema matching, monitoring, etc.
- Current tools are inaccurate
- Extremely high cost to build and maintain systems
  - at enterprises, often at 35% of IT budget [Knoblock et al. 02]
  - hard to build large-scale or long-running systems

The MOBS Approach

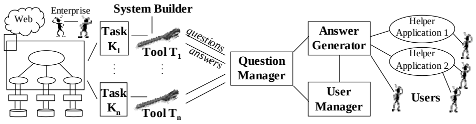
Learn from users to improve tool accuracy, thus significantly reducing builder workload



MOBS = Mass Collaboration to Build Systems

- Many related works employ mass collaboration
  - open-source software, knowledge bases, tech support, software debugging, search engines, recommender systems, ...

MOBS Architecture

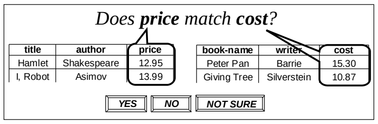


- Answer Generator
  - solicit user answers for User Manager and Question Manager
- User Manager
  - limit user workload and measure user reliabilities
- Question Manager
  - combine user answers to answer questions posed by tools

How to Modify Data Integration Tools

- Ask questions that are hard for automatic tools ...
  - to maximize impact on tool accuracy
- ... but are relatively easy for humans.

Schema Matching Example


Does **price** *match* **cost**?

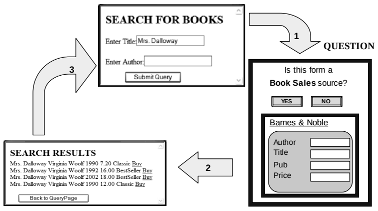
title	author	price
Hamlet	Shakespeare	12.95
I, Robot	Asimov	13.99

book-name	writer	cost
Peter Pan	Barrie	15.30
Giving Tree	Silverstein	10.87

- Currently we ask questions that
  - gather additional training data
  - learn simple domain constraints
  - verify intermediate and final predictions

How to Solicit User Answers

- "Volunteer" settings
  - employees of an organization, online communities
- "Payment" schemes
  - leverage users of existing systems


SEARCH FOR BOOKS

Enter Title: Mrs. Dalloway

Enter Author:

Is this form a **Book Sales** source?

Barnes & Noble

Author
Title
Pub
Price

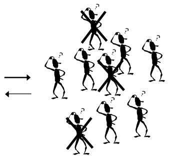
SEARCH RESULTS

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

How to Combine User Answers

- Inject questions with known answers to evaluate users
- Combine answers using user reliability scores
- General framework based on a Dynamic Bayesian Network

Questions for a given task
- those with known answers (User Manager)
- those posed by the tool (Question Manager)



A Working Example of MOBS

(a) MOBS initialization

(b) New questions tq1 and tq2 for tasks k1 and k2; New user Alice

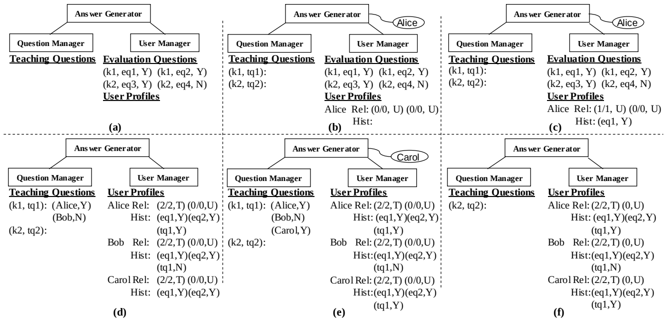
(c) Alice answers correctly on evaluation question for task k1

⋮

(d) Alice, Bob, and Carol are trusted on task k1; Alice and Bob disagreed on question tq1

(e) Carol answers "yes" on tq1

(f) tq1 converges; "yes" is returned to the tool for k1; tq1 is removed



(a)
Teaching Questions
Evaluation Questions
(k1, eq1, Y) (k1, eq2, Y)
(k2, eq3, Y) (k2, eq4, N)
User Profiles

(b)
Teaching Questions
(k1, tq1):
(k2, tq2):
Evaluation Questions
(k1, eq1, Y) (k1, eq2, Y)
(k2, eq3, Y) (k2, eq4, N)
User Profiles
Alice Rel: (0/0, U) (0/0, U)
Hist:

(c)
Teaching Questions
(k1, tq1):
(k2, tq2):
Evaluation Questions
(k1, eq1, Y) (k1, eq2, Y)
(k2, eq3, Y) (k2, eq4, N)
User Profiles
Alice Rel: (1/1, U) (0/0, U)
Hist: (eq1, Y)

(d)
Teaching Questions
(k1, tq1): (Alice,Y)
(Bob,N)
(k2, tq2):
User Profiles
Alice Rel: (2/2,T) (0/0,U)
Hist: (eq1,Y)(eq2,Y)
(tq1,Y)
Bob Rel: (2/2,T) (0/0,U)
Hist: (eq1,Y)(eq2,Y)
(tq1,N)
Carol Rel: (2/2,T) (0/0,U)
Hist: (eq1,Y)(eq2,Y)

(e)
Teaching Questions
(k1, tq1): (Alice,Y)
(Bob,N)
(Carol,Y)
(k2, tq2):
User Profiles
Alice Rel: (2/2,T) (0/0,U)
Hist: (eq1,Y)(eq2,Y)
(tq1,Y)
Bob Rel: (2/2,T) (0/0,U)
Hist:(eq1,Y)(eq2,Y)
(tq1,N)
Carol Rel: (2/2,T) (0/0,U)
Hist:(eq1,Y)(eq2,Y)
(tq1,Y)

(f)
Teaching Questions
(k2, tq2):
User Profiles
Alice Rel: (2/2,T) (0,U)
Hist: (eq1,Y)(eq2,Y)
(tq1,Y)
Bob Rel: (2/2,T) (0,U)
Hist:(eq1,Y)(eq2,Y)
(tq1,N)
Carol Rel: (2/2,T) (0,U)
Hist:(eq1,Y)(eq2,Y)
(tq1,Y)

Empirical Evaluation

Task Types	Domains	Description
Source discovery	Book query interfaces I	24 forms, 17 are bookstore forms
	Faculty directories	30 directories, from 30 departments
1-1 schema matching	Book query interfaces II	10 interfaces, total 65 attributes
	Real estate I	2 schemas, with 55-44 attributes
	Company listings	2 taxonomies, with 330-115 attributes
Complex matching	Real estate II	2 schemas, with 19-32 attributes
	Inventory	2 schemas, with 34-49 attributes

- 3-132 users, used volunteer and payment schemes
- Improved tool accuracy by 9-60%
- Reduced builder workload by 29-88%

Empirical Evaluation

- Users had low workload, answered questions quickly, and their answers were useful
- Extensive simulation confirms previous experiments
  - scaled up to very large populations (tens of thousands)
  - accurate over broad range of population qualities
- Built two simple data integration systems on the Web
  - almost exclusively with user efforts
  - very little builder workload
  - demonstrates potential for building large-scale/long-running systems

Benefits of MOBS

- Frequently the total workload is reduced
  - workload(builder) + workload(users) < workload(builder w/o MOBS)
- Even when total workload increases, can still be very beneficial
  - can speed up the integration process
    - spread workload over multiple users
  - can build systems where not previously possible
    - online communities with members eager to help
  - can enable system expansion
    - free builder to focus on additional improvements

Conclusion & Future Work

- Tools have limited accuracy ⇨ high ownership cost
  - a key bottleneck to widespread deployment of DI systems
- We proposed the MOBS solution
  - make tools learn from multitude of users to improve accuracy
  - ask questions that are easy for humans, hard for machines
- Experiments showed 9-60% accuracy gain, 29-88% workload reduction, and often overall benefits
- Benefits of MOBS
  - speed up integration process
  - build systems where not previously possible
  - free builder to further improve the system

See WebDB-03, TechReport-05 at
http://anhai.cs.uiuc.edu/home/projects/mobs.html