Learning From Multiple Users to Improve Accuracy of Data Integration Tasks


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 [Freelock et al. 02]
  - hard to build large-scale or long-running systems

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.

A Working Example of MOBS

(a) MOBS initialization
(b) New questions t1 and t2 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 disagree on question t1
(e) Carol answers "yes" on t1
(f) t1 converges: "yes" is returned to the tool for k1; t1 is removed

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 or long-running systems

Benefits of MOBS

- Frequently the total workload is reduced
  - workload (builder) < workload (user) + workload (builder w/o MOBS)
- Even when total workload increases, can still be very beneficial
  - can speed up the integration process
  - can build systems where not previously possible
  - can enable system expansion
  - free builders to focus on additional improvements

Conclusion & Future Work

- Tools have limited accuracy (high ownership cost)
- 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 builders to further improve the system

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