Staging User Feedback toward Rapid Conflict Resolution in Data Fusion

Romila Pradhan*, Siarhei Bykau , Sunil Prabhakar*

*Purdue University, Bloomberg L.P.

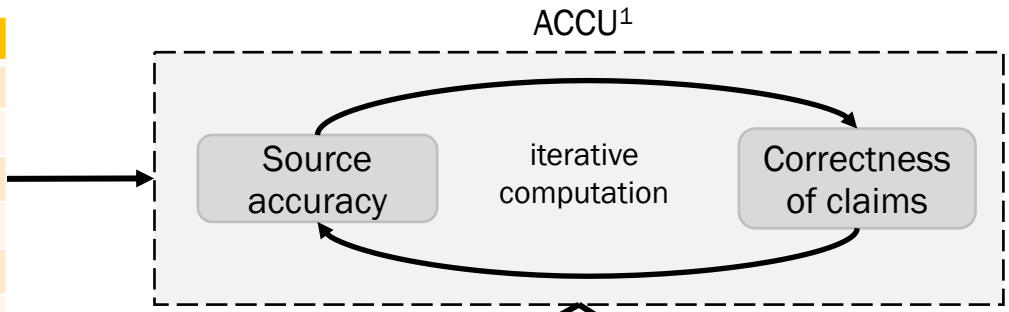
PURDUE UNIVERSITY

Fusing data from multiple sources

Data Item	S_1	S_2	S_3	S_4
Zootopia		Howard	Spencer	Spencer
Kung Fu Panda	Stevenson		Nelson	
Inside Out		leFauve	Docter	
Finding Dory				Stanton
Minions	Coffin	Renaud		
Rio	Jones		Saldanha	

Data fusion systems

Data Item	S_1	S_2	S_3	S_4
Zootopia		Howard	Spencer	Spencer
Kung Fu Panda	Stevenson		Nelson	
Inside Out		leFauve	Docter	
Finding Dory				Stanton
Minions	Coffin	Renaud		
Rio	Jones		Saldanha	



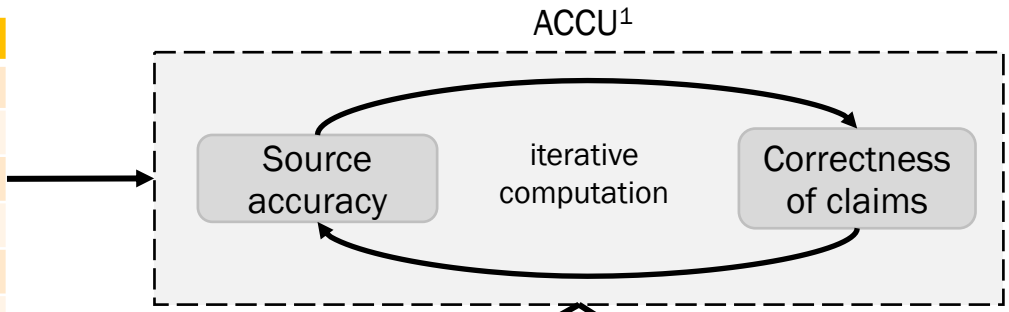
Data Item	Correctness of claims	
Zootopia	Howard (0.000)	Spencer (1.000)
Kung Fu Panda	Stevenson (0.015)	Nelson (0.985)
Inside Out	leFauve (0.001)	Docter (0.999)
Finding Dory	Stanton (1.000)	
Minions	Coffin (0.921)	Renaud (0.079)
Rio	Jones (0.015)	Saldanha (0.985)

Source	Accuracy
S_1	0.317
S_2	0.027
S_3	0.992
S_4	1.000

[1] Xin Luna Dong, Laure Berti-Equille, Divesh Srivastava. Data Fusion: Resolving Conflicts from Multiple Sources. WAIM 2013.

Comparison with ground truth

Data Item	S_1	S_2	S_3	S_4
Zootopia		Howard	Spencer	Spencer
Kung Fu Panda	Stevenson		Nelson	
Inside Out		leFauve	Docter	
Finding Dory				Stanton
Minions	Coffin	Renaud		
Rio	Jones		Saldanha	



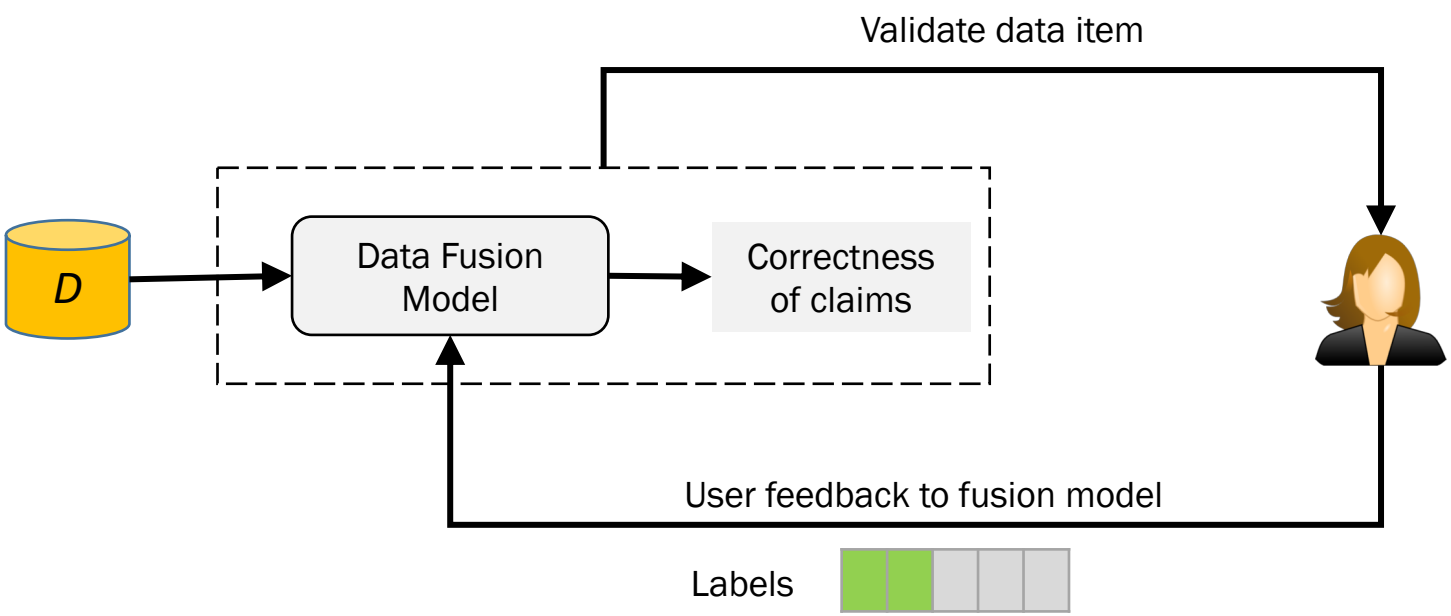
Data Item	Truth
Zootopia	Howard
Kung Fu Panda	Stevenson
Inside Out	Docter
Finding Dory	Stanton
Minions	Coffin
Rio	Saldanha

Data Item	Correctness of claims	
Zootopia	Howard (0.000)	Spencer (1.000)
Kung Fu Panda	Stevenson (0.015)	Nelson (0.985)
Inside Out	leFauve (0.001)	Docter (0.999)
Finding Dory	Stanton (1.000)	
Minions	Coffin (0.921)	Renaud (0.079)
Rio	Jones (0.015)	Saldanha (0.985)

Source	Accuracy
S_1	0.317
S_2	0.027
S_3	0.992
S_4	1.000

[1] Xin Luna Dong, Laure Berti-Equille, Divesh Srivastava. Data Fusion: Resolving Conflicts from Multiple Sources. WAIM 2013.

Involve the User



How to be most effective with user feedback?

This talk

4 ranking strategies

Item-level ranking

Query-by-committee

Uncertainty Sampling

Maximum Expected Utility

Approximate MEU

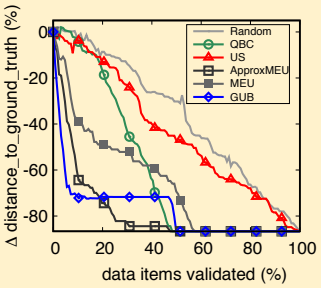
Holistic ranking

Feedback Errors

Non-expert

- Confidence
- Error-rate
- Conflicting feedback

Evaluation



Query-by-committee (QBC)

most sources agree

Data Item	S_1	S_2	S_3	S_4
Zootopia		Howard	Spencer	Spencer
Kung Fu Panda	Stevenson		Nelson	
Inside Out		[illegible]	Docter	
Finding Dory				Stanton
Minions	Coffin	Renaud		
Rio	Jones		Saldanha	

sources disagree

Uncertainty Sampling (US)

Data Item	Correctness of claims	
Zootopia	Howard (0.000)	Spencer (1.000)
Kung Fu Panda	Stevenson (0.015)	Nelson (0.985)
Inside Out	leFauve (0.001)	Docter (0.999)
Finding Dory	Stanton (1.000)	
Minions	Coffin (0.921)	Renaud (0.079)
Rio	Jones (0.015)	Saldanha (0.985)

Implication of a validation

Data Item	S_1	S_2	S_3	S_4
Zootopia		Howard	Spencer	Spencer
Kung Fu Panda	Stevenson		Nelson	
Inside Out		leFauve	Docter	
Finding Dory				Stanton
Minions	Coffin	Renaud		
Rio	Jones		Saldanha	

Implication of a validation

Data Item	S_1	S_2	S_3	S_4
Zootopia		Howard	Spencer	Spencer
Kung Fu Panda	Stevenson		Nelson	
Inside Out		leFauve	Docter	
Finding Dory				Stanton
Minions	Coffin	Renaud		
Rio	Jones		Saldanha	

Ideal utility function

truth function ?

fusion model

data

$$U(D, \mathcal{F}, \mathcal{T}) = \frac{1}{|O|} \left(\sum_{i=1}^{|O|} \sum_{v_i^k \in V_i} p_i^k \delta(\mathcal{T}(v_i^k)) \right)$$

average correctness of true claims

Utility Function

Practical utility function

over correctness of all claims

$$EU(D, \mathcal{F}) = -\sum_{o_i \in O} \sum_{v_i^k \in V_i} p_i^k \log\ p_i^k$$

entropies of all data items

Entropy Utility Function

Maximum Expected Utility (MEU)

- Value of perfect information

entropy utility if claim is true

$$VPI(\theta_i) = EU(D, \mathcal{F}) - \sum_{v_i^k \in V_i} \boxed{EU(D, \mathcal{F} \mid v_i^k = true)} p_i^k$$

Best alternative in the absence of ground truth

Approximate-MEU

- Key idea: Propagation of changes

correctness of claims of validated data item → accuracies of sources → correctness of claims of unvalidated data items

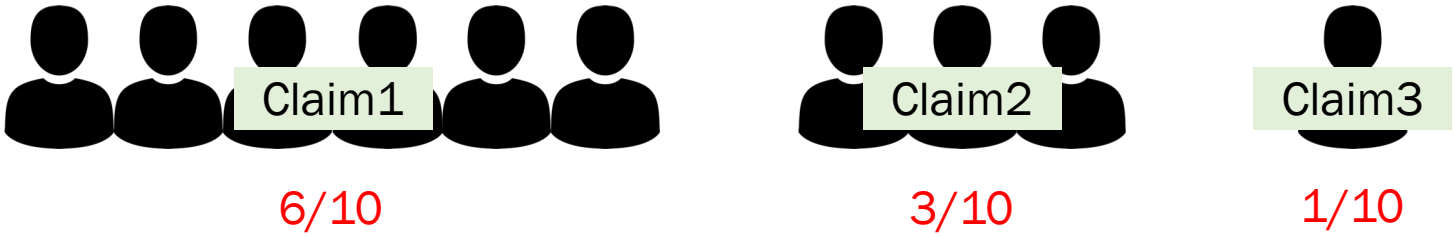
$$VPI(\theta_i) = EU(D, \mathcal{F}) - \sum_{v_i^k \in V_i} EU(D, \mathcal{F} \mid v_i^k = true) p_i^k$$

no need to fuse for every claim!

removed bottleneck iterative computation of MEU

Users can be wrong

- Honest but unsure user — 80% certain about a claim
- Error-rate of user — user is correct 85% of the time
- Conflicting feedback from a crowd of workers



Real-world datasets

	Books[1]	FlightsDay[2]	Population[3]	Flights[2]
Items	1263	5836	40696	121567
Sources	894	38	2545	38
Claims	24303	80452	46734	1931701

Feedback Simulation

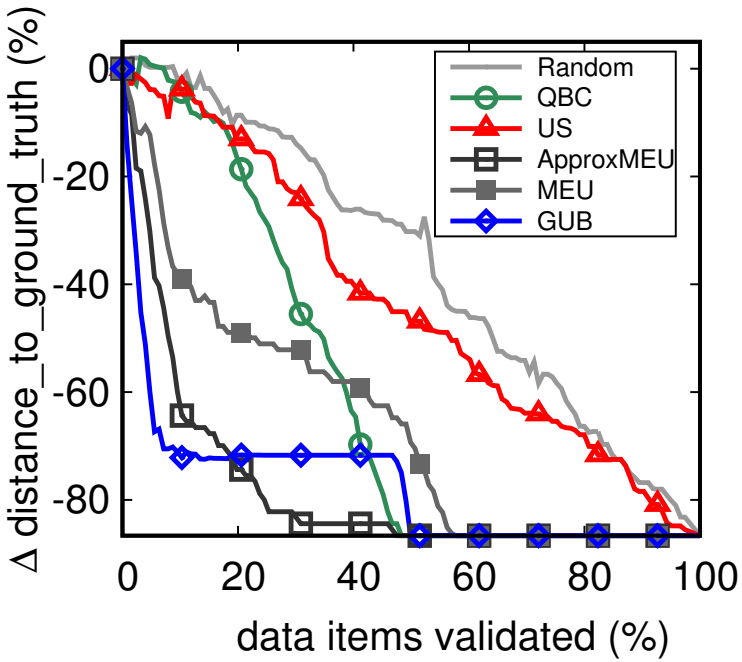
- Books: silver standard provided in [4]
- Flight information: data provided by carrier websites considered ground truth
- Population: manually identified the true claim for data items having multiple claims

1. X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: The role of source dependence. PVLDB, 2009
2. X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava. Truth finding on the deep web: Is the problem solved? PVLDB, 2012
3. J. Pasternack and D. Roth. Knowing what to believe(when you already know something). COLING, 2010
4. http://lunadong.com/fusionDataSets.htm

Competing methods

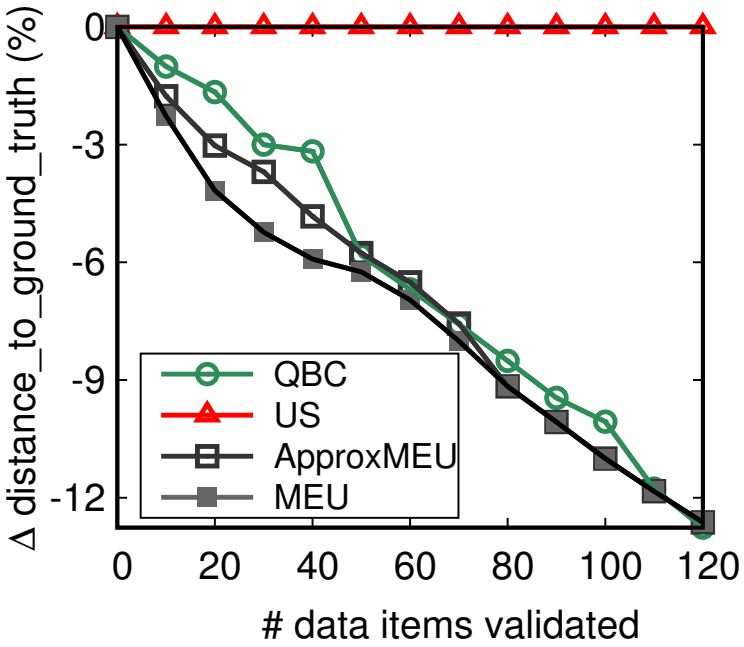
- **Item-level ranking methods**
  - QBC / US
- **Decision-theoretic ranking methods**
  - MEU / Approx-MEU
  - Greedy Upper Bound (GUB) — ground-truth-utility-based
- **Random**
  - all data items equally beneficial

Large number of sources, few claims: holistic ranking



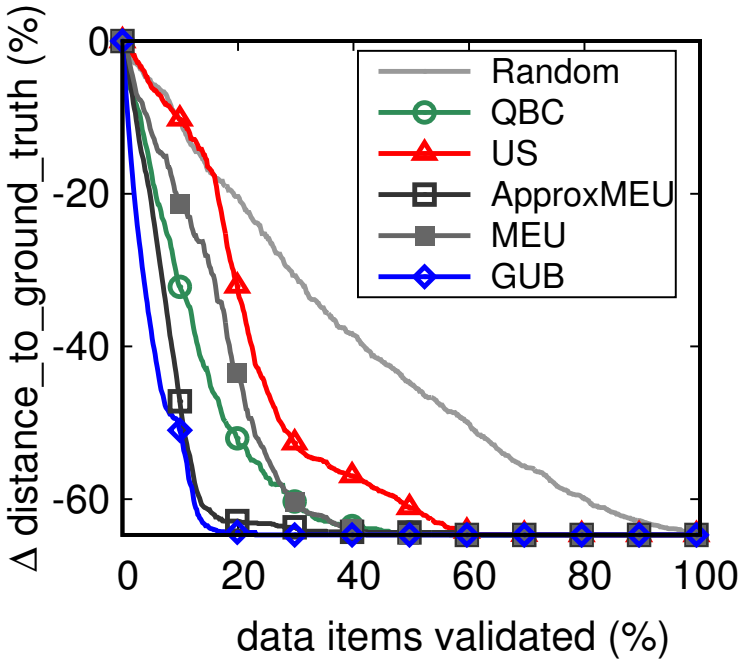
Books

Large number of sources, few claims: holistic ranking



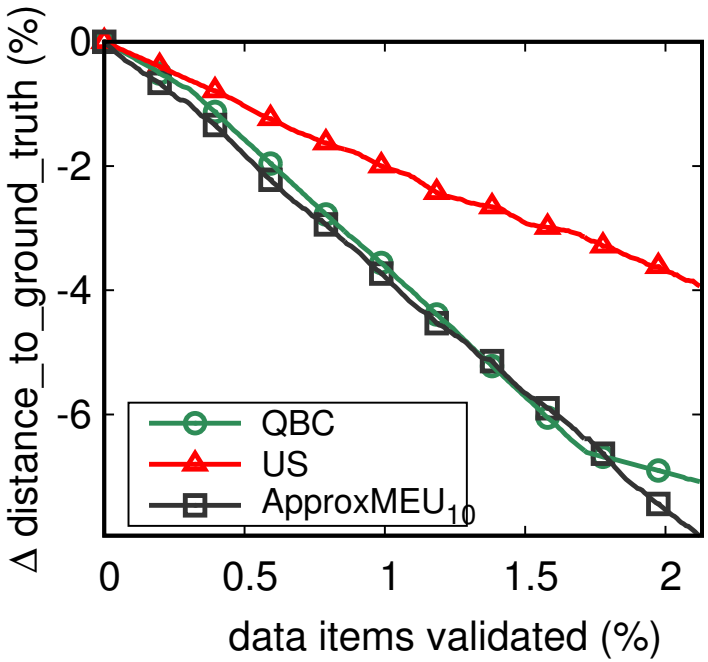
Population

Large number of claims, few sources: either QBC/holistic



FlightsDay

Large number of claims, few sources: either QBC/holistic



Flights

Contributions

- Integrating user feedback to improve the performance of existing data fusion systems
- Designed strategies to generate an effective ordering for validating claims
  - scalable decision-theoretic solution for iterative fusion
  - explored imperfect feedback scenarios
- Evaluation on real-world datasets confirmed that guided feedback rapidly increases the effectiveness of data fusion