Staging User Feedback toward Rapid Conflict Resolution in Data Fusion

Romila Pradhan
Purdue University
rpradhan@cs.purdue.edu

Siarhei Bykau
Bloomberg L.P.
sbykau@bloomberg.net

Sunil Prabhakar
Purdue University
sunil@cs.purdue.edu

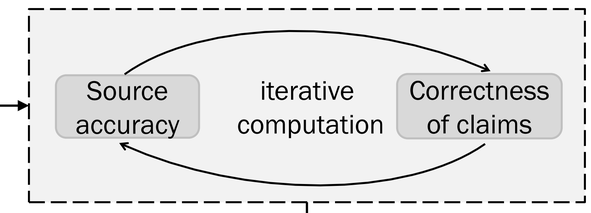
Motivation

- Data fusion systems
  - integrate conflicting data from multiple data sources
  - resolve conflicts to distinguish true/false claims
- Fusion systems, however, are far from perfect:
  - may mislabel a true (false) claim as false (true)

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

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

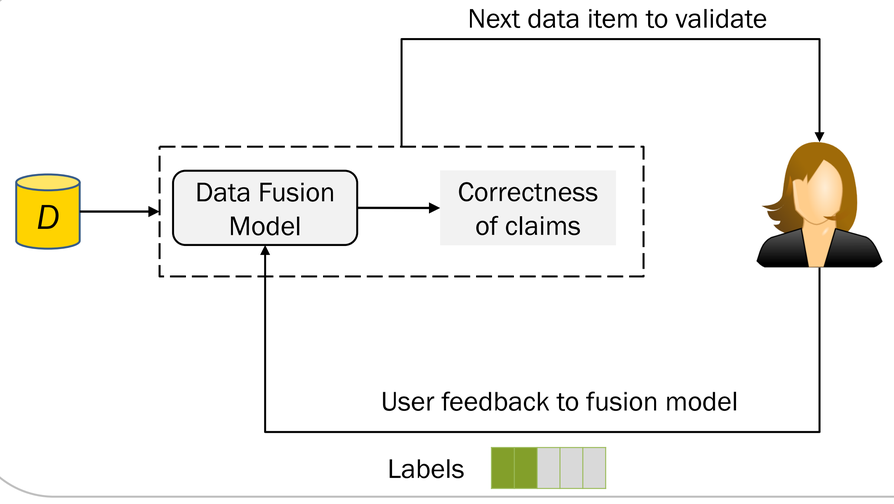
ACCU [Dong et al. WAIM 2013]



Correctness of claims	
Howard (0.000)	Spencer (1.000)
Stevenson (0.015)	Nelson (0.985)
IeFauve (0.001)	Docter (0.999)
Stanton (1.000)	
Coffin (0.921)	Renaud (0.079)
Jones (0.015)	Saldanha (0.985)

claim with higher probability considered correct

Involve the User



- Large number of claims
- Users are expensive; can ask very few questions

How to utilize a user's feedback judiciously?

Ranking Strategies

Query By Committee (QBC)

Data Item	S_1	S_2	S_3	S_4
Zootopia		Howard	Spencer	Spencer
Rio	Jones		Saldanha	

most sources agree about 'Zootopia', disagree about 'Rio'

Uncertainty Sampling (US)

Data Item	Correctness of claims	
Kung Fu Panda	Stevenson (0.015)	Nelson (0.985)
Inside Out	IeFauve (0.001)	Docter (0.999)

fusion model more certain about 'Inside Out' than 'Kung Fu Panda'

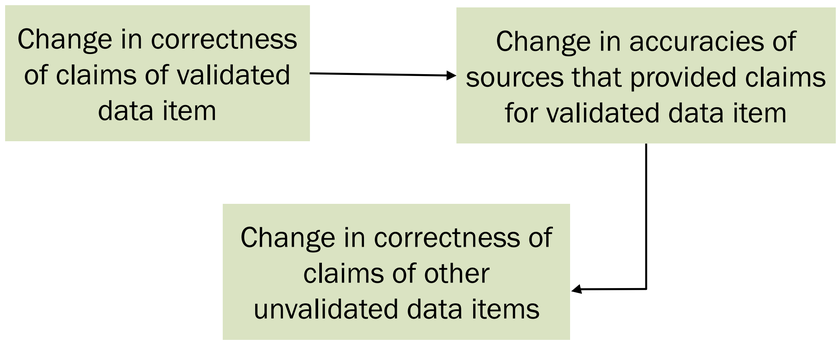
Maximum Expected Utility (MEU)

validating 'Zootopia' can influence more claims than 'Finding Dory'

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

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

Propagation of changes (Approx-MEU)



Feedback Errors

- honest, unsure user
  - e.g., 80% sure about a claim
- error-rate of user
  - e.g., correct 7 out of 10 times
- conflicting feedback

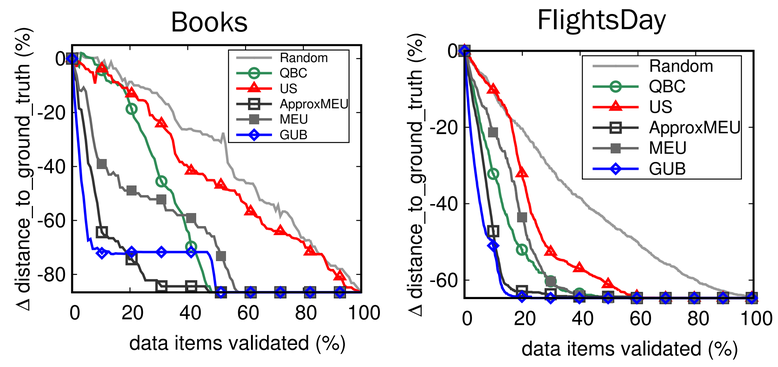
Results

Datasets

- Books [Dong et al. PVLDB 2009]
  - 24Kclaims, 894 sources
- FlightsDay [Li et al. PVLDB 2012]
  - 80K claims, 38 sources
- Population [Pasternack et al. COLING 2010]
  - 47K claims, 2545 sources
- Flights [Li et al. PVLDB 2012]
  - 1.9M claims, 38 sources

Methods

- GUB (ground-truth-based)
- QBC/US/MEU/Approx-MEU our methods



- Guided feedback improves the performance of fusion
  - MEU performs better on long-tail data
  - QBC comparable on dense data
- Complete set of results in paper

Summary

- Proposed judicious use of user feedback to improve the performance of existing data fusion systems
- Designed strategies to generate an effective ordering for validating claims
  - can scale decision-theoretic solution to iterative fusion models using propagation of changes
  - explored imperfect feedback scenarios