Corroborating Information from Disagreeing Views

Alban Galland[1] Serge Abiteboul[1]
Amélie Marian[2] Pierre Senellart[3]

[1] INRIA Saclay–Île-de-France [2] Rutgers University [3] Télécom ParisTech

February 4, 2010, *WSDM*

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SACLAY – ÎLE-DE-FRANCE

Motivating Example

What are the capital cities of European countries?

	France	Italy	Poland	Romania	Hungary
Alice	Paris	Rome	Warsaw	Bucharest	Budapest
Bob	?	Rome	Warsaw	Bucharest	Budapest
Charlie	Paris	Rome	Katowice	Bucharest	Budapest
David	Paris	Rome	Bratislava	Budapest	Sofia
Eve	Paris	Florence	Warsaw	Budapest	Sofia
Fred	Rome	?	?	Budapest	Sofia
George	Rome	?	?	?	Sofia

Voting

Information: redundance

	France	Italy	Poland	Romania	Hungary
Alice	Paris	Rome	Warsaw	Bucharest	Budapest
Bob	?	Rome	Warsaw	Bucharest	Budapest
Charlie	Paris	Rome	Katowice	Bucharest	Budapest
David	Paris	Rome	Bratislava	Budapest	Sofia
Eve	Paris	Florence	Warsaw	Budapest	Sofia
Fred	Rome	?	?	Budapest	Sofia
George	Rome	?	?	?	Sofia
Frequence	**P.** 0.67 R. 0.33	**R.** 0.80 F. 0.20	**W.** 0.60 K. 0.20 B. 0.20	Buch. 0.50 Bud. 0.50	Bud. 0.43 **S.** 0.57

Evaluating Trustworthiness of Sources

Information: redundance, trustworthiness of sources (= average frequence of predicted correctness)

	France	Italy	Poland	Romania	Hungary	Trust
Alice	Paris	Rome	Warsaw	Bucharest	Budapest	0.60
Bob	?	Rome	Warsaw	Bucharest	Budapest	0.58
Charlie	Paris	Rome	Katowice	Bucharest	Budapest	0.52
David	Paris	Rome	Bratislava	Budapest	Sofia	0.55
Eve	Paris	Florence	Warsaw	Budapest	Sofia	0.51
Fred	Rome	?	?	Budapest	Sofia	0.47
George	Rome	?	?	?	Sofia	0.45
Frequence	**P.** 0.70	**R.** 0.82	**W.** 0.61	**Buch.** 0.53	Bud. 0.46	
weighted	R. 0.30	F. 0.18	K. 0.19	Bud. 0.47	**S.** 0.54	
by trust			B 0.20			

Iterative Fixpoint Computation

Information: redundance, trustworthiness of sources with iterative fixpoint computation

	France	Italy	Poland	Romania	Hungary	Trust
Alice	Paris	Rome	Warsaw	Bucharest	Budapest	0.65
Bob	?	Rome	Warsaw	Bucharest	Budapest	0.63
Charlie	Paris	Rome	Katowice	Bucharest	Budapest	0.57
David	Paris	Rome	Bratislava	Budapest	Sofia	0.54
Eve	Paris	Florence	Warsaw	Budapest	Sofia	0.49
Fred	Rome	?	?	Budapest	Sofia	0.39
George	Rome	?	?	?	Sofia	0.37
Frequence	**P.** 0.75	**R.** 0.83	**W.** 0.62	**Buch.** 0.57	**Bud.** 0.51	
weighted	R. 0.25	F. 0.17	K. 0.20	Bud. 0.43	S. 0.49	
by trust			B 0.19			

Context and problem

- Context:
  - Set of sources stating facts
  - (Possible) functional dependencies between facts
  - Fully unsupervised setting: we do not assume any information on truth values of facts or inherent trust in sources
- Problem: determine which facts are true and which facts are false
- Real world applications: query answering, source selection, data quality assessment on the web, making good use of the wisdom of crowds

Outline

Introduction

Model

Algorithms

Experiments

Conclusion

Outline

Introduction

Model

Algorithms

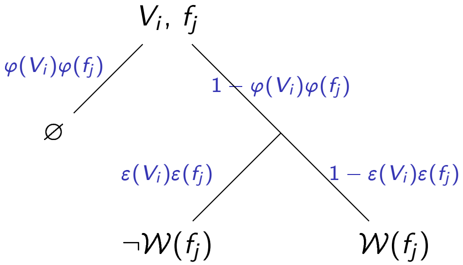
Experiments

Conclusion

General Model

- Set of facts $\mathcal{F} = \{f_1 \ldots f_n\}$
  - Examples: "Paris is capital of France", "Rome is capital of France", "Rome is capital of Italy"
- Set of views (= sources) $\mathcal{V} = \{V_1 \ldots V_m\}$, where a view is a partial mapping from $\mathcal{F}$ to {T, F}
  - Example:
    ¬ "Paris is capital of France" ∧ "Rome is capital of France"
- Objective: find the most likely real world $\mathcal{W}$ given $\mathcal{V}$ where the real world is a total mapping from $\mathcal{F}$ to {T, F}
  - Example:
    "Paris is capital of France" ∧ ¬ "Rome is capital of France" ∧
    "Rome is capital of Italy" ∧ ...

Generative Probabilistic Model



$V_i,\ f_j$
$\varphi(V_i)\varphi(f_j)$
$1-\varphi(V_i)\varphi(f_j)$
$\varnothing$
$\varepsilon(V_i)\varepsilon(f_j)$
$1-\varepsilon(V_i)\varepsilon(f_j)$
$\neg\mathcal{W}(f_j)$
$\mathcal{W}(f_j)$

- $\varphi(V_i)\varphi(f_j)$: probability that V_i “forgets” f_j
- $\varepsilon(V_i)\varepsilon(f_j)$: probability that V_i “makes an error” on f_j
- Number of parameters: $n+2(n+m)$
- Size of data: $\tilde{\varphi}nm$ with $\tilde{\varphi}$ the average forget rate

Obvious Approach

- Method: use this generative model to find the most likely parameters given the data
  - Inverse the generative model to compute the probability of a set of parameters given the data
- Not practically applicable:
  - Non-linearity of the model and boolean parameter $\mathcal{W}(f_j)$
    ⇒ equations for inversing the generative model very complex
  - Large number of parameters (n and m can both be quite large)
    ⇒ Any exponential technique unpractical

⇒ Heuristic fix-point algorithms

Outline

Introduction

Model

Algorithms

Experiments

Conclusion

Baselines

Counting (does not look at negative statements, popularity)

$$\begin{cases} T & \text{if } \dfrac{\left|\{V_i : V_i(f_j) = T\}\right|}{\max_f \left|\{V_i : V_i(f) = T\}\right|} \geqslant \eta \\ F & \text{otherwise} \end{cases}$$

Voting (adapted only with negative statements)

$$\begin{cases} T & \text{if } \dfrac{\left|\{V_i : V_i(f_j) = T\}\right|}{\left|\{V_i : V_i(f_j) = T \vee V_i(f_j) = F\}\right|} \geqslant 0.5 \\ F & \text{otherwise} \end{cases}$$

TruthFinder [YHY07]: heuristic fix-point method from the literature

3-Estimates

- Iterative estimation of 3 kind of parameters:
  - truth value of facts
  - error rate or trustworthiness of sources
  - hardness of facts
- Tricky normalization to ensure stability

Functional dependencies

- So far, the models and algorithms are about positive and negative statements, without correlation between facts
- How to deal with functional dependencies (e.g., capital cities)?

pre-filtering: When a view states a value, all other values governed by this FD are considered stated false. *If I say that Paris is the capital of France, then I say that neither Rome nor Lyon nor . . . is the capital of France.*

post-filtering: Choose the best answer for a given FD.

Outline

Introduction

Model

Algorithms

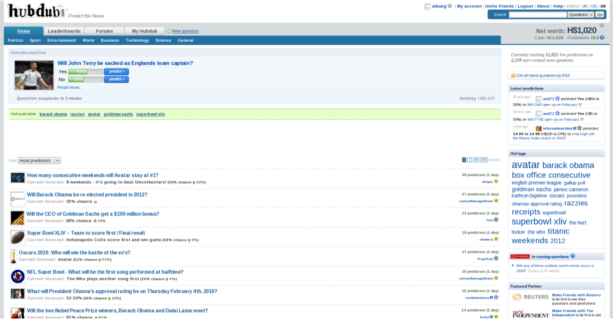
Experiments

Conclusion

Datasets

- Synthetic dataset: large scale and higly customizable
- Real-world datasets:
    - General-knowledge quiz
    - Biology 6th-grade test
    - Search-engines results
    - Hubdub

Hubdub (1/2)



Net worth: H$1,020
Cash: H$ 1,020 Predictions: H$ 0

FEATURED QUESTION

Will John Terry be sacked as Englands team captain?

Yes 55%

No 45%

Read more...

Question suspends in 3 weeks

Activity: H$6,923

POPULAR NOW: barack obama razzies avatar goldman sachs superbowl xliv

How many consecutive weekends will Avatar stay at #1?
Current forecast: 8 weekends - it's going to beat Ghostbusters! (59% chance 59%)
39 predictors (1 day)
shojun

Will Barack Obama be re-elected president in 2012?
Current forecast: 35% chance
27 predictors (1 day)
carnacthemagnificent

Will the CEO of Goldman Sachs get a $100 million bonus?
Current forecast: 38% chance 19%
22 predictors (1 day)
lola

Super Bowl XLIV – Team to score first / Final result
Current forecast: Indianapolis Colts score first and win game (58% chance 8%)
19 predictors (1 day)
chatarra

Oscars 2010: Who will win the battle of the ex's?
Current forecast: Avatar (51% chance 51%)
17 predictors (1 day)
frogchop

NFL Super Bowl - What will be the first song performed at halftime?
Current forecast: The Who plays another song first (24% chance 4%)
16 predictors (1 day)
carnacthemagnificent

What will President Obama's approval rating be on Thursday February 4th, 2010?
Current forecast: 52-53% (36% chance 20%)
15 predictors (1 day)
notablenotices

Will the two Nobel Peace Prize winners, Barack Obama and Dalai Lama meet?
Current forecast: 81% chance 81%
14 predictors (1 day)
kruijs

*Currently tracking **51,821** live predictions on **2,219** user-created news questions*

Get all latest questions by RSS

Latest predictions

31 secs ago aed71 predicted **Yes** (H$50 at 50%) on Will DAX open up on February 3?

55 secs ago aed71 predicted **Yes** (H$5 at 50%) on Will FTSE open up on February 3?

1 min ago infernalmachine predicted **14.00 to 14.99** (H$100 at 24%) on How high will the Misery Index reach in 2010?

Hot tags

avatar barack obama box office consecutive english premier league gallup poll goldman sachs james cameron kathryn bigelow oscars president obamas approval rating razzies receipts superbowl superbowl xliv the hurt locker the who titanic weekends 2012

In-running In-running questions

- Will any of these unlikely world events occur in 2010? Closes in 47 weeks

Featured Partner

REUTERS Make Friends with Reuters to be first to see their questions and predictions.

THE INDEPENDENT Make Friends with The Independent to be first to see

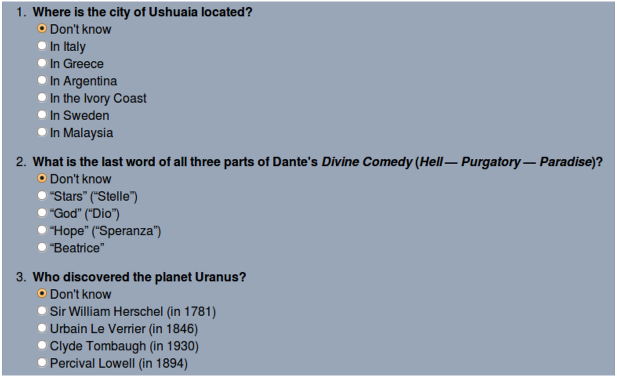
`http://www.hubdub.com/`

- 357 questions, 1 to 20 answers, 473 participants

Hubdub (2/2)

	Number of errors (no post-filtering)	Number of errors (with post-filtering)
Voting	278	292
Counting	340	327
TruthFinder	458	274
3-Estimates	272	270

General-Knowledge Quiz (1/2)



1. **Where is the city of Ushuaia located?**
   - ● Don't know
   - ○ In Italy
   - ○ In Greece
   - ○ In Argentina
   - ○ In the Ivory Coast
   - ○ In Sweden
   - ○ In Malaysia
2. **What is the last word of all three parts of Dante's *Divine Comedy* (*Hell* — *Purgatory* — *Paradise*)?**
   - ● Don't know
   - ○ “Stars” (“Stelle”)
   - ○ “God” (“Dio”)
   - ○ “Hope” (“Speranza”)
   - ○ “Beatrice”
3. **Who discovered the planet Uranus?**
   - ● Don't know
   - ○ Sir William Herschel (in 1781)
   - ○ Urbain Le Verrier (in 1846)
   - ○ Clyde Tombaugh (in 1930)
   - ○ Percival Lowell (in 1894)

`http://www.madore.org/~david/quizz/quizz1.html`

- 17 questions, 4 to 14 answers, 601 participants

General-Knowledge Quiz (2/2)

	Number of errors (no post-filtering)	Number of errors (with post-filtering)
Voting	11	6
Counting	12	6
TruthFinder	-	-
3-Estimates	9	0

Outline

Introduction

Model

Algorithms

Experiments

Conclusion

In brief

- We believe truth discovery is an important problem, we do not claim we have solved it completely
- Collection of fix-point methods (see paper), one of them (3-Estimates) performing remarkably and regularly well
- Cool real-world applications!

All code and datasets available from
`http://datacorrob.gforge.inria.fr/`

Thanks.

Webdam

Foundations of Web data management

Perspectives

- Exploiting dependencies between sources [DBES09]
- Numerical values (1.77m and 1.78m cannot be seen as two completely contradictory statements for a height)
- No clear functional dependencies, but a limited number of values for a given object (e.g., phone numbers)
- Pre-existing trust, e.g., in a social network
- Clustering of facts, each source being trustworthy for a given field

References I

- Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava.
Integrating conflicting data: The role of source dependence.
In *Proc. VLDB*, Lyon, France, August 2009.

- Xiaoxin Yin, Jiawei Han, and Philip S. Yu.
Truth discovery with multiple conflicting information providers on the Web.
In *Proc. KDD*, San Jose, California, USA, August 2007.