Corroborating Information from Disagreeing Views

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

1 INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE INRIA centre de recherche SACLAY - ÎLE-DE-FRANCE 2 RUTGERS 3 TELECOM ParisTech

May 10, 2010, *DBWeb meeting*

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
Decision	Paris	Rome	Warsaw	?	Sofia

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 weighted by trust	P. 0.70 R. 0.30	R. 0.82 F. 0.18	W. 0.61 K. 0.19 B 0.20	Buch. 0.53 Bud. 0.47	Bud. 0.46 S. 0.54	
Decision	Paris	Rome	Warsaw	Bucharest	Sofia	

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 weighted by trust	P. 0.75 R. 0.25	R. 0.83 F. 0.17	W. 0.62 K. 0.20 B 0.19	Buch. 0.57 Bud. 0.43	Bud. 0.51 S. 0.49	
Decision	Paris	Rome	Warsaw	Bucharest	Budapest	

Some Complications

- There might be no explicit contradictions between facts stated by different sources:
  - "Paris is a city of France."
  - "Lyon is a city of France."
  - "Bolzano is a city of France."
  - $\neg$ "New York is a city of France."
- We want to exploit the fact that some facts are harder than other (capital of France vs capital of Vanuatu).

Context and problem

- Context:
  - Set of sources stating facts
  - (Possible) functional dependencies between facts
  - Fully unsupervised setting: we do not assume any information on the truth values of facts or the inherent trust of 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

1. Introduction
2. Model
3. Algorithms
4. Experiments
5. Conclusion

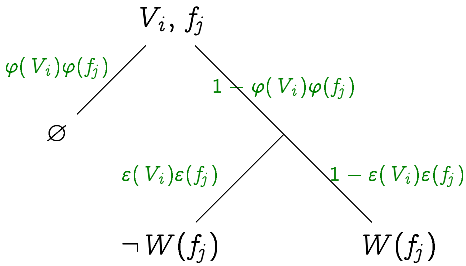
Outline

1. Introduction
2. Model
3. Algorithms
4. Experiments
5. 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:
    $\neg$ "Paris is capital of France" $\wedge$ "Rome is capital of France"
- Objective: find the most likely real world W given $\mathcal{V}$ where the real world is a total mapping from $\mathcal{F}$ to {T, F}
  - Example:
    "Paris is capital of France" $\wedge$ $\neg$ "Rome is capital of France" $\wedge$
    "Rome is capital of Italy"

Generative Probabilistic Model



- $\varphi(V_i)\varphi(f_j)$: probability that V_i "forgets" (does not state anything about) f_j
- $\varepsilon(V_i)\varepsilon(f_j)$: probability that V_i makes an error on f_j if V_i makes a statement about f_j
- Number of parameters: $n + 2(n + m)$ (n boolean parameters, $2(n + m)$ parameters between 0 and 1).
- 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
  - Standard machine learning technique: Expectation-Maximization
- Not practically applicable:
  - Equations for inversing the generative model very complex (but doable)
  - Large number of parameters (n and m can both be quite large). Any exponential technique unpractical
  - Non-linearity of the model ($W(f_j)$ is boolean)
- $\Rightarrow$ Heuristic fix-point algorithms

PageRank

- PageRank [BP98]: Fix-point algorithm for computing authority scores on the Web
- Corresponds to the equilibrium measure of the random walk in the (slightly modified) Web graph
- Can it be applied directly?
  - Sources-Facts: bipartite graph. Random walks (obviously) do not converge in this setting.
  - Alternative: Graph of the two-steps paths in this bipartite graph. Random walks work, but it can be shown that the equilibrium measure is proportional to the degree (cf. method Counting further)
  - No clear notion how to manage negative statements (negative links)
- Source of inspiration for the methods presented

Outline

1. Introduction
2. Model
3. Algorithms
4. Experiments
5. Conclusion

Baselines

Counting (does not look at negative statements, popularity)

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

Voting (adapted only with negative statements)

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

TruthFinder [YHY07]: heuristic fix-point method from the literature; context slightly different (Source-Object-Fact) and method most adapted to cases with very few errors, does not deal with contradiction

Fix-Point Algorithms

1. Estimate the truth of facts (e.g., with voting)
2. Based on that, estimate the error rates of sources
3. Based on that, refine the estimation for the facts
4. Based on that, refine the estimation for the sources
5. ...

Iterate until a fix-point is reached (and cross your fingers it converges!).

Cosine

- The truth of a fact is what views state weighted by how error prone they are
- The error of a view is the correlation (= cosine similarity) between its statement of facts and the predicted truth of these facts

Precise algorithms are given in [GAMS10].

2-Estimates

- A fact is true:
  - if a view states it is true and makes no error
  - or if a view states it is false and makes an error
- A view makes an error:
  - if it states a fact is true and the fact is false
  - if it states a fact is false and the fact is true
- Quite instable $\Rightarrow$ tricky normalization

3-Estimates

- Similar in spirit to 2-Estimates but estimation of 3 parameters:
  - truth value of facts
  - error rate or trustworthiness of sources
  - hardness of facts
- Also needs tricky normalization

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

1 Introduction

2 Model

3 Algorithms

4 Experiments

5 Conclusion

Experiments: Generalities

What to measure?

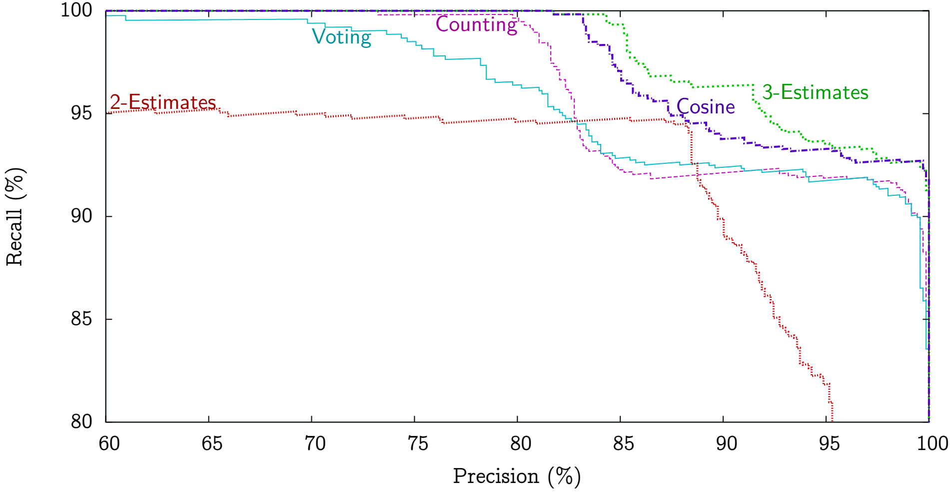
- Quality of binary classification: percentage of error for predicting the truth
- Precision-Recall curve for top-k rated facts (classical measure for search engine results)

On what data?

- Synthetic dataset closely based upon our generative model, with all possibilities of variation
- Various real-world datasets

We assume that error rates are less than 50%!

Typical Results over Synthetic Dataset



Hubdub: 1/2

http://www.hubdub.com/

- Prediction network (sports, politics, business, etc.)
- Bets using virtual money
- (Small) sports dataset extracted: 357 questions, 1 to 20 answers, 473 users, 3,051 statements (before pre-filtering)

Hubdub: 2/2

	Number of errors (no post-filtering)	Number of errors (with post-filtering)
Voting	278	292
Counting	340	327
TruthFinder	458	274
2-Estimates	269	269
Cosine	357	357
3-Estimates	272	270

Hubdub: 2/2

	Number of errors (no post-filtering)	Number of errors (with post-filtering)
Voting	278	292
Counting	340	327
TruthFinder	458	274
2-Estimates	269	269
Cosine	357	357
3-Estimates	272	270

Possible to earn money on bets. Easy way to get rich!

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	-	-
2-Estimates	6	6
Cosine	7	6
3-Estimates	9	0

General-Knowledge Quiz: 2/2

	Number of errors (no post-filtering)	Number of errors (with post-filtering)
Voting	11	6
Counting	12	6
TruthFinder	-	-
2-Estimates	6	6
Cosine	7	6
3-Estimates	9	0

Possible to know the correct answer to a quiz by just looking at all answers. Automatic correction of exams is possible!

It does not always work!

No magic!

- Does not take into account dependencies between sources
- Example: integration of search engine results
- Usually, when it "does not work", 3-Estimates gives results comparable to the baseline, Cosine is not bad, 2-Estimates is very unstable

Outline

1 Introduction

2 Model

3 Algorithms

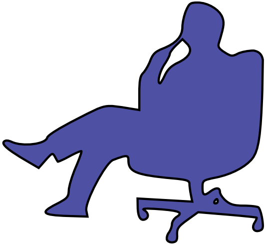
4 Experiments

5 Conclusion

In brief

- One of the first works in truth discovery among disagreeing sources
- Collection of fix-point methods, one of them (3-Estimates) performing remarkably and regularly well
- We believe this is an important problem, we do not claim we have solved it completely
- Cool real-world applications!

All code and datasets available from `http://datacorrob.gforge.inria.fr/`. Details in [GAMS10].



- 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

Merci.

Foundations of Web data management

References I

Sergey Brin and Lawrence Page.
The anatomy of a large-scale hypertextual Web search engine.
Computer Networks and ISDN Systems, 30(1–7):107–117, 1998.

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

Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart.
Corroborating information from disagreeing views.
In *Proc. WSDM*, pages 1041–1064, New York, USA, February 2010.

References II

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.