Towards Veracity Challenge in Big Data

Jing Gao\textsuperscript{1}, Qi Li\textsuperscript{1}, Bo Zhao\textsuperscript{2}, Wei Fan\textsuperscript{3}, and Jiawei Han\textsuperscript{4}

\textsuperscript{1}SUNY Buffalo; \textsuperscript{2}LinkedIn;
\textsuperscript{3}Baidu Research Big Data Lab; \textsuperscript{4}University of Illinois
Big data challenge

• **Volume**
  • The quantity of generated and stored data
Big data challenge

• **Velocity**
  • The speed at which the data is generated and processed
Big data challenge

• **Variety**
  • The type and nature of the data
Big data challenge

• **Veracity**
  • The quality of captured data
Causes of Veracity Issue

• Rumors
• Spammers
• Collection errors
• Entry errors
• System errors
• ...

Aspects of Solving Veracity Problems

• **Sources and claims**
  • We know who claims what
  • Truth discovery

• **Features of sources and claims**
  • Features of sources, eg. history, graphs of sources
  • Features of claims, eg. hashtags, lexical patterns
  • Rumor detection
  • Source trustworthiness analysis
Overview

1. Introduction
2. Truth Discovery: Veracity Analysis from Sources and Claims
3. Truth Discovery Scenarios
4. Veracity Analysis from Features of Sources and Claims
5. Applications
6. Open Questions and Resources
7. References
Truth Discovery

• **Problem**

  • Input: Multiple conflicting information about the same set of objects provided by various information sources
  
  • Goal: Discover trustworthy information (i.e., the **truths**) from conflicting data on the same object
Example 1: Knowledge Base Construction

• Knowledge base
  – Construct knowledge base based on huge amount of information on Internet

• Problem
  – Find true facts from multiple conflicting sources
What Is The Height Of Mount Everest?
**Mount Everest - Wikipedia, the free encyclopedia**

By the same measure of base to summit, Mount McKinley, in Alaska, is also taller than Everest. Despite its height above sea level of only 6,193.6 m (20,320 ft),...

List of deaths on eight ... - Edmund Hillary - Timeline of climbing Mount - 1996

---

**Mt Everest Height Mystery May Be Answered : Discovery News**

Mar 1, 2012 – The plunge from 71581 feet was a success. Next up: 120000 feet.

---

**Facts About Mt Everest**

Number of people to successfully climb Mt. Everest 660. Number of people who have died trying to climb Mt. Everest: 142. Height: 29,020 feet, or 5 and a half ...

---

**Mount Everest by the Numbers: Deaths, Cost to Climb, and More ...**

May 22, 2012

8,000: Height in meters (approximately 26,000 feet) at Mount Everest's “death zone,” the low-oxygen area above ...

More videos for what is the height of mount everest »

---

**What is the height of Mount Everest**

Mt. Everest is 29,022 feet high. And 348,024 inches high. What is the real height of Mount Everest? 12,000 ft!!! Everest is, to begin with, 18,000 ft above sea level ...

---

**Height of Mount Everest (Everest, Mount) -- Britannica Online ...**

The height of Mount Everest, according to the most recent and reliable data, is 29035 feet (8850 metres). In 1999 an American survey, sponsored by the (U.S.) ...

---

**Mount Everest - Overview of Mount Everest**

With a peak elevation of 29,035 feet (8850 meters), the top of Mount Everest is the world's highest point above sea level. As the world's highest mountain, ...
Mount Everest - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Mount_Everest
By the same measure of base to summit, Everest. Despite its height above sea level, List of deaths on eight ... - Edmund Hillary

Mt. Everest Height: Mystery May E news.discovery.com/.../everest-official-height-Mar 1, 2012 – The plunge from 71581 feet

Facts About Mt. Everest
teacher.scholastic.com/activities/hillary/arm Number of people to successfully climb Mount Everest died trying to climb Mt. Everest: 142. Hei

Mount Everest by the Numbers: D www.thedailybeast.com/May 22, 2012 8,000: Height in meters of Everest’s “death zone”

More videos for what is the height of mount everest

What is the height of Mount Everest
wiki.answers.com: Geography > Land Mt. Everest is 29,022 feet high. And 348 Mount Everest is 22,000 ft!!! Everest is, to

Height of Mount Everest (Everest
www.britannica.com/EBchecked/.../Height
The height of Mount Everest, according to feet (8850 metres). In 1999 an American

Mount Everest - Overview of Mount
geography.about.com: Specific Place With a peak elevation of 29,035 feet (8848) world's highest point above sea level. As the world's highest mountain, ...
Which of these square numbers also happens to be the sum of two smaller square numbers?

- A: 16
- B: 25
- C: 36
- D: 49

Example 2: Crowdsourced Question Answering

https://www.youtube.com/watch?v=BbX44YSsQ2I
Aggregation
A Straightforward Aggregation Solution

- **Voting/Averaging**
  - Take the value that is claimed by majority of the sources
  - Or compute the mean of all the claims

- **Limitation**
  - Ignore source reliability

- **Source reliability**
  - Is crucial for finding the true fact but unknown
Which of these square numbers also happens to be the sum of two smaller square numbers?

A: 16  
B: 25  
C: 36  
D: 49

https://www.youtube.com/watch?v=BbX44YSsQ2I
A Straightforward Aggregation Solution

• Voting/Averaging
  – Take the value that is claimed by majority of the sources
  – Or compute the mean of all the claims

• Limitation
  – Ignore source reliability

• Source reliability
  – Is crucial for finding the true fact but unknown
Truth Discovery

• Principle
  – Infer both truth and source reliability from the data
    • A source is reliable if it provides many pieces of true information
    • A piece of information is likely to be true if it is provided by many reliable sources
Model Categories

• Optimization model (OPT)
• Statistical model (STA)
• Probabilistic graphical model (PGM)
Optimization Model (OPT)

• General model

\[
\arg \min_{\{w_s, \{v_o^*\}\}} \sum \sum g(w_s, v_o^*) \\
\text{s.t. } \delta_1(w_s) = 1, \delta_2(v_o^*) = 1
\]

• What does the model mean?
  • Find the optimal solution that minimize the objective function
  • Jointly estimate true claims $v_o^*$ and source reliability $w_s$ under some constraints $\delta_1, \delta_2, \ldots$.
  • Function $g(\cdot,\cdot)$ can be distance, entropy, etc.
Optimization Model (OPT)

• General model

\[
\arg \min_{\{w_s\}, \{v_o^*\}} \sum_{o \in O} \sum_{s \in S} g(w_s, v_o^*) \\
\text{s. t. } \delta_1(w_s) = 1, \delta_2(v_o^*) = 1
\]

• How to solve the problem?
  • Use the method of Lagrange multipliers
  • Block coordinate descent to update parameters
  • If each sub-problem is convex and smooth, then convergence is guaranteed
OPT - CRH Framework

\[
\begin{align*}
\min_{X^(*), \mathcal{W}} \quad f(X^(*), \mathcal{W}) &= \sum_{k=1}^{K} w_k \sum_{i=1}^{N} \sum_{m=1}^{M} d_m (v_{im}^{(*)}, v_{im}^{(k)}) \\
\text{s.t.} \quad \delta(\mathcal{W}) &= 1, \quad \mathcal{W} \geq 0.
\end{align*}
\]

CRH is a framework that deals with the heterogeneity of data. Different data types are considered, and the estimation of source reliability is jointly performed across all the data types together.

[Li et al., SIGMOD’14]
**OPT - CRH Framework**

\[
\min_{x^{(*)}, \mathcal{W}} f(x^{(*)}, \mathcal{W}) = \sum_{k=1}^{K} w_k \sum_{i=1}^{N} \sum_{m=1}^{M} d_m (v_{im}^{(*)}, v_{im}^{(k)})
\]

s.t. \( \delta(\mathcal{W}) = 1, \quad \mathcal{W} \geq 0. \)

**Basic idea**

- Truths should be close to the claims from reliable sources
- Minimize the overall weighted distance to the truths in which reliable sources have high weights
OPT - CRH Framework

• **Loss function**
  - $d_m$: loss on the data type of the $m$-th property
  - Output a high score when the claim deviates from the truth
  - Output a low score when the claim is close to the truth

• **Constraint function**
  - The objective function may go to $-\infty$ without constraints
  - Regularize the weight distribution
OPT - CRH Framework

• Run the following until convergence

  • Truth computation
    • Minimize the weighted distance between the truth and the sources’ claims

\[
\nu_{im}^{(\ast)} \leftarrow \arg \min_{\nu} \sum_{k=1}^{K} w_k \cdot d_m \left( \nu, \nu_{im}^{(k)} \right)
\]

  • Source reliability estimation
    • Assign a weight to each source based on the difference between the truths and the claims made by the source

\[
\mathcal{W} \leftarrow \arg \min_{\mathcal{W}} f \left( \mathcal{X}^{(\ast)}, \mathcal{W} \right)
\]
Statistical Model (STA)

• General goal:
  - To find the (conditional) probability of a claim being true

• Source reliability:
  - Probability(ies) of a source/worker making a true claim
Statistical Model (STA)

• Models
  ➢ Apollo-MLE [Wang et al., ToSN’14]
  ➢ TruthFinder [Yin et al., TKDE’08]
  ➢ Investment, Pool Investment [Pasternack&Roth, COLING’10]
  ➢ Cosine, 2-estimate, 3-estimate [Galland et al., WSDM’10]
Different websites often provide conflicting information on a subject, e.g., Authors of “Rapid Contextual Design”

<table>
<thead>
<tr>
<th>Online Store</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Powell’s books</td>
<td>Holtzblatt, Karen</td>
</tr>
<tr>
<td>Barnes &amp; Noble</td>
<td>Karen Holtzblatt, Jessamyn Wendell, Shelley Wood</td>
</tr>
<tr>
<td>A1 Books</td>
<td>Karen Holtzblatt, Jessamyn Burns Wendell, Shelley Wood</td>
</tr>
<tr>
<td>Cornwall books</td>
<td>Holtzblatt-Karen, Wendell-Jessamyn Burns, Wood</td>
</tr>
<tr>
<td>Mellon’s books</td>
<td>Wendell, Jessamyn</td>
</tr>
<tr>
<td>Lakeside books</td>
<td>WENDELL, JESSAMYNHOLTZBLATT, KARENWOOD, SHELLEY</td>
</tr>
<tr>
<td>Blackwell online</td>
<td>Wendell, Jessamyn, Holtzblatt, Karen, Wood, Shelley</td>
</tr>
</tbody>
</table>

[Yin et al., TKDE’08]
Each object has a set of conflicting facts
  - E.g., different author lists for a book
And each web site provides some facts
How to find the true fact for each object?

STA - TruthFinder
1. There is usually only one true fact for a property of an object
2. This true fact appears to be the same or similar on different web sites
   • E.g., “Jennifer Widom” vs. “J. Widom”
3. The false facts on different web sites are less likely to be the same or similar
   • False facts are often introduced by random factors
4. A web site that provides mostly true facts for many objects will likely provide true facts for other objects
Confidence of facts ↔ Trustworthiness of web sites

- A fact has *high confidence* if it is provided by (many) trustworthy web sites
- A web site is *trustworthy* if it provides many facts with high confidence

Iterative steps

- Initially, each web site is equally trustworthy
- Based on the four heuristics, infer fact confidence from web site trustworthiness, and then backwards
- Repeat until achieving stable state
STA - TruthFinder

Web sites

\[ w_1 \]
\[ w_2 \]
\[ w_3 \]
\[ w_4 \]

Facts

\[ f_1 \]
\[ f_2 \]
\[ f_3 \]
\[ f_4 \]

Objects

\[ o_1 \]
\[ o_2 \]
STA - TruthFinder

Web sites: $w_1, w_2, w_3, w_4$

Facts: $f_1, f_2, f_3, f_4$

Objects: $o_1, o_2$
STA - TruthFinder

Web sites

\( w_1 \)  \( f_1 \)
\( w_2 \)  \( f_2 \)
\( w_3 \)  \( f_3 \)
\( w_4 \)  \( f_4 \)

Facts

Objects

\( o_1 \)
\( o_2 \)
STA - TruthFinder

Web sites

- $w_1$
- $w_2$
- $w_3$
- $w_4$

Facts

- $f_1$
- $f_2$
- $f_3$
- $f_4$

Objects

- $o_1$
- $o_2$
STA - TruthFinder

• The trustworthiness of a web site \( w \): \( t(w) \)
  - Average confidence of facts it provides

\[
t(w) = \frac{\sum_{f \in F(w)} s(f)}{|F(w)|}
\]

- Sum of fact confidence
- Set of facts provided by \( w \)

• The confidence of a fact \( f \): \( s(f) \)
  - One minus the probability that all web sites providing \( f \) are wrong

\[
s(f) = 1 - \prod_{w \in W(f)} (1 - t(w))
\]

- Probability that \( w \) is wrong
- Set of websites providing \( f \)
Probabilistic Graphical Model (PGM)

Source reliability

Truth
Probabilistic Graphical Model (PGM)

- **Models**
  - GTM [Zhao & Han, QDB’12]
  - LTM [Zhao et al., VLDB’12]
  - MSS [Qi et al., WWW’13]
  - LCA [Pasternack & Roth, WWW’13]
  - TEM [Zhi et al., KDD’15]
  ...

PGM – Gaussian Truth Model (GTM)

- **Real-valued** Truths and Claims
  - Population of a city is numerical
- The quality of sources is modeled as how close their claims are to the truth
  - Distance is better than accuracy for numerical data
- Sources and objects are **independent** respectively

[Zhao&Han, QDB’12]
PGM – Gaussian Truth Model (GTM)

Quality of Sources

Claims

Truth of Facts
**PGM – Gaussian Truth Model (GTM)**

- For each source $k$
  - Generate its quality from a prior inverse Gamma distribution: 
    \[ \sigma^2_s \sim Inv - Gamma(\alpha, \beta) \]
- For each fact $f$
  - Generate its prior truth from a prior Gaussian distribution: 
    \[ \mu_e \sim Gaussian(\mu_0, \sigma^2_0) \]
- For each claim $c$ of fact $f$, generate claim of $c$.
  - Generate it from a Gaussian distribution with truth as mean and the quality as variance: 
    \[ o_c \sim Gaussian(\mu_e, \sigma^2_{sc}) \]
Overview

1. Introduction
2. Truth Discovery: Veracity Analysis from Sources and Claims
3. Truth Discovery Scenarios
4. Veracity Analysis from Features of Sources and Claims
5. Applications
6. Open Questions and Resources
7. References
Number of Truths for One Object

• **Single truth**
  • Each object has one and only one truth
  • The claims from sources contain the truth
  • Complementary vote

• **Multiple truth**
  • Each object may have more than true fact
  • Each source may provide more than one fact for each object

• **Existence of truths**
  • The true fact for an object may be not presented by any sources
Single Truth

• **Example**
  • A person’s birthday
  • Population of a city
  • Address of a shop

• **Complementary vote**
  • If a source makes a claim on an object, that source considers all the other claims as false

• **Positive vote only** [Wang et al., ToSN’14]
  • An event only receive positive claims, but no negative claims. E.g., people only report that they observe an event.
Multiple Truth- Latent Truth Model (LTM)

- **Multiple** facts can be **true** for each entity (object)
  - One book may have 2+ authors
- A source can make **multiple claims per entity**, where more than one of them can be true
  - A source may claim a book w. 3 authors
- **Source reliability**
  - False positive: making a wrong claim
  - Sensitivity: missing a claim
- Modeled in PGM

[Zhao et al., VLDB’12]
Multiple Truth- Latent Truth Model (LTM)

False positive rate
sensitivity

Truth of Facts
Multiple Truth- Latent Truth Model (LTM)

• For each source $k$
  • Generate false positive rate (with strong regularization, believing most sources have low FPR): $\phi_k^0 \sim \text{Beta}(\alpha_{0,1}, \alpha_{0,0})$
  • Generate its sensitivity (1-FNR) with uniform prior, indicating low FNR is more likely: $\phi_k^1 \sim \text{Beta}(\alpha_{1,1}, \alpha_{1,0})$

• For each fact $f$
  • Generate its prior truth prob, uniform prior: $\theta_f \sim \text{Beta}(\beta_1, \beta_0)$
  • Generate its truth label: $t_f \sim \text{Bernoulli}(\theta_f)$

• For each claim $c$ of fact $f$, generate claim of $c$.
  • If $f$ is false, use false positive rate of source: $o_c \sim \text{Bernoulli}(\phi_{sc}^0)$
  • If $f$ is true, use sensitivity of source: $o_c \sim \text{Bernoulli}(\phi_{sc}^1)$
Existence of Truth

- **Truth Existence problem**: when the true answers are excluded from the candidate answers provided by all sources.
  - *Has-truth questions*: correct answers exist among the candidate answers provided by all sources.
  - *No-truth questions*: true answers are not included in the candidate answers provided by all sources.
- Without any prior knowledge, the no-truth questions are hard to distinguish from the has-truth ones.
  - These no-truth questions degrade the precision of the answer integration system.
- **Example: Slot Filling Task**

[Yu et al., COLING’14][Zhi et al., KDD’15]
## Existence of Truth

### Example: Slot Filling Task

Table 1: Example Questions of Slot Filling Task

<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q_1$</td>
</tr>
<tr>
<td>$q_2$</td>
</tr>
<tr>
<td>$q_3$</td>
</tr>
<tr>
<td>$q_4$</td>
</tr>
<tr>
<td>$q_5$</td>
</tr>
<tr>
<td>$q_6$</td>
</tr>
<tr>
<td>$q_7$</td>
</tr>
<tr>
<td>$q_8$</td>
</tr>
</tbody>
</table>

Stuart Rose

**Businessman**

Stuart Alan Ransom Rose, Baron Rose of Monowden is a British businessman, who was the executive chairman of the British retailer Marks & Spencer. For this role he was paid an annual salary of £1,130,000. [Wikipedia](https://en.wikipedia.org/wiki/Stuart_Rose)

**Born:** March 17, 1949 (age 65), Gosport, United Kingdom

**Education:** Bootham School
## Existence of Truth

<table>
<thead>
<tr>
<th>Source</th>
<th>q1</th>
<th>q2</th>
<th>q3</th>
<th>q4</th>
<th>q5</th>
<th>q6</th>
<th>q7</th>
<th>q8</th>
</tr>
</thead>
<tbody>
<tr>
<td>s₁</td>
<td>43</td>
<td></td>
<td></td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s₂</td>
<td>:11</td>
<td></td>
<td>:31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>UK</td>
</tr>
<tr>
<td>s₃</td>
<td>627</td>
<td>Pakistan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s₄</td>
<td></td>
<td>Afghanistan</td>
<td>Ghazni</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s₅</td>
<td>43</td>
<td></td>
<td></td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s₆</td>
<td>43</td>
<td></td>
<td></td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s₇</td>
<td>43</td>
<td>Afghanistan</td>
<td>Ghazni</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s₈</td>
<td>IL</td>
<td>Khost</td>
<td>Mar</td>
<td>London</td>
<td>Marks</td>
<td>Russia</td>
<td>Holly</td>
<td></td>
</tr>
<tr>
<td>s₉</td>
<td>/25</td>
<td>Pakistan</td>
<td>Pakistan</td>
<td></td>
<td>actor</td>
<td>Spence</td>
<td>Spence</td>
<td></td>
</tr>
<tr>
<td>s₁₀</td>
<td></td>
<td></td>
<td></td>
<td>Kabul</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s₁₁</td>
<td>43</td>
<td></td>
<td></td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s₁₂</td>
<td></td>
<td>Afghanistan</td>
<td>Ghazni</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s₁₃</td>
<td>9</td>
<td></td>
<td></td>
<td>50</td>
<td>Holland</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Truth</td>
<td>43</td>
<td>Afghanistan</td>
<td>Ghazni</td>
<td>50</td>
<td>Empty</td>
<td>Empty</td>
<td>Empty</td>
<td>Empty</td>
</tr>
</tbody>
</table>

- **Has-truth questions**
- **No-truth questions**
Existence of Truth - Truth Existence Model (TEM)

• Probabilistic Graphical Model
  • Output
    • $t$: latent truths
    • $\phi$: source quality
  • Input
    • $A$: observed answers
    • $S$: sources
  • Parameters (fixed)
    • Prior of source quality: $\alpha$
    • Prior of truth: $\eta$  
      \[
      \eta_{i0} = P(t_i = E), \eta_{in} = P(t_i = d_{in})
      \]
• Maximum Likelihood Estimation
• Inference: EM
Many truth discovery methods consider independent sources.

- Sources provide information independently.
- Source correlation can be hard to model.
- However, this assumption may be violated in real life.

Copy relationships between sources:

- Sources can copy information from one or more other sources.

General correlations of sources.
Known relationships

- Apollo-Social [Wang et al., IPSN’14]
  - For a claim, a source may copy from a related source with a certain probability
  - Used MLE to estimate a claim being correct

Unknown relationships

- Accu-Copy [Dong et al., VLDB’09a] [Dong et al., VLDB’09b]
- MSS [Qi et al., WWW’13]
  - Modeled as a PGM
  - Related sources are grouped together and assigned with a group weight
Copy Relationships between Sources

• High-level intuitions for copying detection
  • Common error implies copying relation
    • e.g., many same errors in $s_1 \cap s_2$ imply source 1 and 2 are related
  • Source reliability inconsistency implies copy direction
    • e.g., $s_1 \cap s_2$ and $s_1 - s_2$ has similar accuracy, but $s_1 \cap s_2$ and $s_2 - s_1$ has different accuracy, so source 2 may be a copier.
Incorporate copying detection in truth discovery

Step 2

Truth Discovery

Step 3

Source-accuracy Computation

Step 1

Copying Detection

[Dong et al., VLDB’09a] [Dong et al., VLDB’09b]
General Source Correlation

• More general source correlations
  • Sources may provide data from complementary domains (negative correlation)
  • Sources may focus on different types of information (negative correlation)
  • Sources may apply common rules in extraction (positive correlation)

• How to detect
  • Hypothesis test of independence using joint precision and joint recall

[Pochampally et al., SIGMOD’14]
Information Density

• Dense information
  • Each source provides plenty of claims
  • Each object receives plenty of information from sources

• Long-tail phenomenon on sources side
  • Many sources provide limited information
  • Only a few sources provide sufficient information

• Auxiliary information
  • Text of question/answers
  • Fine-grained source reliability estimation
Long-tail Phenomenon on Sources Side

![Graph showing a power law function fit for the number of sources versus the number of claims. The graph includes a power law function fit line.](graph.png)
Long-tail Phenomenon on Sources Side - CATD

• Challenge when most sources make a few claims
  • Sources weights are usually estimated as proportional to the accuracy of the sources
  • If long-tail phenomenon occurs, most source weights are not properly estimated.

• A confidence-aware approach
  • not only estimates source reliability
  • but also considers the confidence interval of the estimation

• An optimization based approach

[Li et al., VLDB’15]
Long-tail Phenomenon on Sources Side - CATD

• Assume that sources are independent and error made by source $s$: $\epsilon_s \sim N(0, \sigma_s^2)$

• $\epsilon_{aggregate} = \frac{\sum_{s \in S} w_s \epsilon_s}{\sum_{s \in S} w_s} \sim N\left(0, \frac{\sum_{s \in S} w_s^2 \sigma_s^2}{(\sum_{s \in S} w_s)^2}\right)$

Without loss of generality, we constrain $\sum_{s \in S} w_s = 1$

• Optimization

\[
\min_{\{w_s\}} \sum_{s \in S} w_s^2 \sigma_s^2
\]

s.t. $\sum_{s \in S} w_s = 1,$

$w_s \geq 0, \forall s \in S.$
Long-tail Phenomenon on Sources Side - CATD

Sample variance:

$$\hat{\sigma}_s^2 = \frac{1}{|N_s|} \sum_{n \in N_s} (x_n^s - x_n^{*(0)})^2$$

where $x_n^{*(0)}$ is the initial truth.

The estimation is not accurate with small number of samples.

Find a range of values that can act as good estimates.

Calculate confidence interval based on

$$\frac{|N_s| \hat{\sigma}_s^2}{\sigma_s^2} \sim \chi^2(|N_s|)$$
• Consider the possibly worst scenario of $\sigma_s^2$
• Use the upper bound of the 95% confidence interval of $\sigma_s^2$

$$u_s^2 = \sum_{n \in N_s} \left( x_n^s - x_n^* (0) \right)^2$$

$$\frac{\chi^2}{\chi(0.05, |N_s|)}$$
Long-tail Phenomenon on Sources Side - CATD

\[
\begin{align*}
\min_{\{w_s\}} & \quad \sum_{s \in S} w_s^2 u_s^2 \\
\text{s.t.} & \quad \sum_{s \in S} w_s = 1, w_s \geq 0, \forall s \in S.
\end{align*}
\]

• Closed-form solution:

\[
w_s \propto \frac{1}{u_s^2} = \frac{\chi^2_{(0.05, |N_s|)}}{\sum_{n \in N_s} \left( x_n^s - x_n^* (0) \right)^2}
\]
Example on calculating confidence interval

<table>
<thead>
<tr>
<th>Source ID</th>
<th># Claims</th>
<th>$\hat{\sigma}^2$</th>
<th>Confidence Interval (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source A</td>
<td>200</td>
<td>0.1</td>
<td>(0.0830, 0.1229)</td>
</tr>
<tr>
<td>Source B</td>
<td>200</td>
<td>3</td>
<td>(2.4890, 3.6871)</td>
</tr>
<tr>
<td>Source C</td>
<td>2</td>
<td>0.1</td>
<td>(0.0271, 3.9498)</td>
</tr>
<tr>
<td>Source D</td>
<td>2</td>
<td>3</td>
<td>(0.8133, 118.49)</td>
</tr>
</tbody>
</table>
## Long-tail Phenomenon on Sources Side - CATD

### Example on calculating source weight

<table>
<thead>
<tr>
<th>Source ID</th>
<th>( \hat{\sigma}_s^2 )</th>
<th>( u_s^2 )</th>
<th>Source Weight (based on ( \hat{\sigma}_s^2 ))</th>
<th>Source Weight (based on ( u_s^2 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source A</td>
<td>0.1</td>
<td>0.1229</td>
<td>0.4839</td>
<td>0.9385</td>
</tr>
<tr>
<td>Source B</td>
<td>3</td>
<td>3.6871</td>
<td>0.0161</td>
<td>0.0313</td>
</tr>
<tr>
<td>Source C</td>
<td>0.1</td>
<td>3.9498</td>
<td>0.4839</td>
<td>0.0292</td>
</tr>
<tr>
<td>Source D</td>
<td>3</td>
<td>118.49</td>
<td>0.0161</td>
<td>0.0010</td>
</tr>
</tbody>
</table>
Long-tail Phenomenon on Sources Side - CATD

<table>
<thead>
<tr>
<th>Question level</th>
<th>Error rate of Majority Voting</th>
<th>Error rate of CATD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0297</td>
<td>0.0132</td>
</tr>
<tr>
<td>2</td>
<td>0.0305</td>
<td>0.0271</td>
</tr>
<tr>
<td>3</td>
<td>0.0414</td>
<td>0.0276</td>
</tr>
<tr>
<td>4</td>
<td>0.0507</td>
<td>0.0290</td>
</tr>
<tr>
<td>5</td>
<td>0.0672</td>
<td>0.0435</td>
</tr>
<tr>
<td>6</td>
<td>0.1101</td>
<td>0.0596</td>
</tr>
<tr>
<td>7</td>
<td>0.1016</td>
<td>0.0481</td>
</tr>
<tr>
<td>8</td>
<td>0.3043</td>
<td>0.1304</td>
</tr>
<tr>
<td>9</td>
<td>0.3737</td>
<td>0.1414</td>
</tr>
<tr>
<td>10</td>
<td>0.5227</td>
<td>0.2045</td>
</tr>
</tbody>
</table>

Higher level indicates harder questions
Fine-Grained Truth Discovery - FaitCrowd

• To learn fine-grained (topical-level) user expertise and the truths from conflicting crowd-contributed answers.
• Topic is learned from question&answer texts

[Ma et al., KDD’15]
Fine-Grained Truth Discovery - FaitCrowd

- **Input**
  - Question Set
  - User Set
  - Answer Set
  - Question Content

- **Output**
  - Questions’ Topic
  - Topical-Level Users’ Expertise
  - Truths
Jointly modeling question content and users’ answers by introducing latent topics.

Modeling question content can help estimate reasonable user reliability, and in turn, modeling answers leads to the discovery of meaningful topics.

Learning topics, topic-level user expertise and truths simultaneously.
**Fine-Grained Truth Discovery - FaitCrowd**

- **Answer Generation**
  - The correctness of a user’s answer may be affected by the question’s topic, user’s expertise on the topic and the question’s bias.
  - Draw user’s expertise
    \[ e_{z_u} \sim N(\mu, \sigma^2) \]
• Answer Generation

The correctness of a user’s answer may be affected by the question’s topic, user’s expertise on the topic and the question’s bias.

• Draw user’s expertise
  \( e_{z_q u} \sim N(\mu, \sigma^2) \)

• Draw the truth
  \( t_q \sim U(\gamma_q) \)
Fine-Grained Truth Discovery - FaitCrowd

• Answer Generation
  • The correctness of a user’s answer may be affected by the question’s topic, user’s expertise on the topic and the question’s bias.
    • Draw user’s expertise
      \[ e_{z_q u} \sim N(\mu, \sigma^2) \]
    • Draw the truth
      \[ t_q \sim U(\gamma_q) \]
    • Draw the bias
      \[ b_q \sim N(0, \sigma^{2'}) \]
**Fine-Grained Truth Discovery - FaitCrowd**

**Answer Generation**

- The correctness of a user’s answer may be affected by the question’s topic, user’s expertise on the topic and the question’s bias.
  - Draw user’s expertise
    \[ e_{zq}u \sim N(\mu, \sigma^2) \]
  - Draw the truth
    \[ t_q \sim U(\gamma_q) \]
  - Draw the bias
    \[ b_q \sim N(0, \sigma^{2'}) \]
  - Draw a user’s answer
    \[ a_{qu}|t_q \sim \text{logistic}(e_{zq}u, b_q) \]

\[ e_{zq}u \uparrow \text{ and } b_q \downarrow \longrightarrow p(a_{qu} = t_q|t_q) \uparrow \]
\[ e_{zq}u \downarrow \text{ and } b_q \uparrow \longrightarrow p(a_{qu} = t_q|t_q) \downarrow \]
## Fine-Grained Truth Discovery - FaitCrowd

<table>
<thead>
<tr>
<th>Question level</th>
<th>Majority Voting</th>
<th>CATD</th>
<th>FaitCrowd</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0297</td>
<td>0.0132</td>
<td>0.0132</td>
</tr>
<tr>
<td>2</td>
<td>0.0305</td>
<td>0.0271</td>
<td>0.0271</td>
</tr>
<tr>
<td>3</td>
<td>0.0414</td>
<td>0.0276</td>
<td>0.0241</td>
</tr>
<tr>
<td>4</td>
<td>0.0507</td>
<td>0.0290</td>
<td>0.0254</td>
</tr>
<tr>
<td>5</td>
<td>0.0672</td>
<td>0.0435</td>
<td>0.0395</td>
</tr>
<tr>
<td>6</td>
<td>0.1101</td>
<td>0.0596</td>
<td>0.0550</td>
</tr>
<tr>
<td>7</td>
<td>0.1016</td>
<td>0.0481</td>
<td>0.0481</td>
</tr>
<tr>
<td>8</td>
<td>0.3043</td>
<td>0.1304</td>
<td>0.0870</td>
</tr>
<tr>
<td>9</td>
<td>0.3737</td>
<td>0.1414</td>
<td>0.1010</td>
</tr>
<tr>
<td>10</td>
<td>0.5227</td>
<td>0.2045</td>
<td>0.1136</td>
</tr>
</tbody>
</table>
Source reliability evolves over time

Update source reliability based on continuously arriving data:

\[ p(w_s|e_{1:T}^s) \propto p(e_T^s|w_s)p(w_s|e_{1:T-1}^s) \]

[Li et al., KDD’15]
Overview

1. Introduction
2. Truth Discovery: Veracity Analysis from Sources and Claims
3. Truth Discovery Scenarios
4. Veracity Analysis from Features of Sources and Claims
5. Applications
6. Open Questions and Resources
7. References
Veracity Analysis from Features of Sources and Claims

- Rumor detection
  - Find the rumor
  - Find the source of the rumor
- Source trustworthiness analysis
  - Graph based model
  - Learning based model
Rumor Detection on Twitter

- Clues for Detecting Rumors
  - Burst
  - High retweet ratio
  - Clue words

Fig. 7. Diagram of process flow

[Takahashi&Igata, SCIS’12]
Rumor Detection – Find the Rumor

• Content-based features
  • Lexical patterns
  • Part-of-speech patterns

• Network-based features
  • Tweeting and retweeting history

• Microblog-specific memes
  • Hashtags
  • URLs
  • Mentions

[Qazvinian et al., EMNLP’11][Ratkiewicz et al., CoRR’10]
Rumor Detection on Sina Weibo

• Content-based features
  • Has multimedia, sentiment, has URL, time span

• Network-based features
  • Is retweeted, number of comments, number of retweets

• Client
  • Client program used

• Account
  • Gender of user, number of followers, user name type, ...

• Location
  • Event location

[Yang et al., MDS’12]
Rumor Detection – Find the Source

• Graph G
• If u infected, v not, and u-v, u will infect v after delay $\sim \exp(\lambda)$
• Note: everyone will be infected, just a matter of time.

[Shah&Zaman, SIGMETRICS’12]
Centrality Measures

• How “important” or central is a node $u$?

• Rank or measure with topological properties
  • Degree
  • Eigenvector
  • Pagerank
  • Betweenness
    • The fraction of all shortest paths that a node $u$ is on
  • Closeness
    • Average of shortest distances from $u$ to other nodes
    • Equal to rumor centrality for trees

Fig. 7. A network where the distance center does not equal the general graph rumor center.
Rumor Source Detection – Rumor Centrality

- Known infinite regular tree $G$, degree $d > 1$
- $\exp(\lambda)$ transmission times
  - Each edge has iid random draw
  - Value is the same for either direction
- At an unknown time $t$, you observe the state of the network.
- Which node was the source of the infection?
- Idea: Compute rumor centrality for each node in infected subgraph; take highest ranking node
Rumor Source Detection – Rumor Suspects

• Here you also have an a priori set of suspects $S$

• Which suspect was the source of the infection?

• Idea: Compute rumor centrality like before, but take highest ranking node in $S$

[Dong et al., ISIT’13]
Rumor Source Detection – Multiple Observations

• Here you have multiple observations of independent rumor spreads, with the same source.

• Idea: Compute rumor centrality for each graph, take product

\[
\hat{s} := \arg \max_{s \in \bigcap_{i=1}^{m} G_i} \prod_{i=1}^{m} R(s, G_i)
\]

[Wang et al., SIGMETRICS’14]
Source Trustworthiness — Graph-Based

• Intuition
  • A page has a high trustworthiness if its backlinks are trustworthy

• Only use source linkage

[Page et al., 1999] [Kleinberg, JACM’99]
Source Trustworthiness – EigenTrust

• Problem in P2P:
  • Inauthentic files distributed by malicious nodes

• Objective:
  • Identify the source of inauthentic files and bias against downloading from them

• Basic Idea
  • Each peer has a *Global Reputation* given by the local trust values assigned by other peers

[Kamvar et al., WWW’03]
Source Trustworthiness – EigenTrust

• Local trust value $c_{ij}$
  • The opinion peer $i$ has of peer $j$, based on past experiences
  • Each time peer $i$ downloads an authentic/inauthentic file from peer $j$, $c_{ij}$ increases/decreases.

• Global trust value $t_i$
  • The trust that the entire system places in peer $i$

$$t_{ik} = \sum_j c_{ij} c_{jk}$$

What their opinion of peer $k$

Ask friend $j$

Weight your friend’s opinion by how much you trust them
Source Trustworthiness – Learning-Based

• Trust prediction: classification problem
  • Trust: positive class
  • Not trust: negative class

• Features
  • Extracted from sources to represent pairs of users
Developed extensive list of possible predictive variables for trust between users

- User factors
- Interaction factors

Epinions
- Write reviews
- Rate reviews
- Post comments

Used several ML tools
- Decision tree
- Naïve Bayes
- SVM
- Logistic regression

Interaction factors are important to predict trust

Source Trustworthiness – User Pair Trust

[Figure 2: An taxonomy of user factors]

[Figure 3: An taxonomy of interaction factors]

[Liu et al., EC’08]
Overview

1. Introduction
2. Truth Discovery: Veracity Analysis from Sources and Claims
3. Truth Discovery Scenarios
4. Veracity Analysis from Features of Sources and Claims
5. Applications
6. Open Questions and Resources
7. References
Applications

• Knowledge base construction
  • Slot filling

• Social media data analysis
  • Rumor/fraud detection, rumor propagation
  • Claim aggregation

• Mobile sensing
  • Environmental monitoring

• Wisdom of the crowd
  • Community question answering systems
Mobile Sensing

Human Sensor
PM2.5 value?
Health-Oriented Community Question Answering Systems
My husband had quintuple heart bypass surgery one year ago today. He is experiencing increasing amount of pain in his left leg where a vein was removed. He describes it as a squeezing pain below the knee; like his leg will be squeezed in half at times. Doctors don't seem to be able to find the cause, indicating that it may be nerve damage. Pain management medications don't always seem to ease the pain either. Anyone else experiencing this or does anyone have any insight? Thanks!

Tags: leg pain, Heart surgery
Quality of Question-Answer Thread

By nikgonz | Jan 18, 2008

20 Comments

AntiqueLady001 | Jun 12, 2013
To: TerrySa

I'm 76 and have a similar problem with the leg they took the vein from and it's been 6 years since my triple bypass. The heart doctor just said my circulation was less than it should be in my legs but offered no solution or medication for the pain. I think the pain comes from my lower back. I never had back problems or leg pain before this surgery. I did find that a muscle relaxant like Valium worked to stop the nerve pain at night and allowed me to sleep. But like anything else simple that works, after 9 months my primary doctor refused to refill my prescription.

I only took one-half of a 5 mg tablet a night so I don't know why he kept trying to prescribe meds that were stronger and had side effects. But he did and now I refuse to take a med that once I start it I can't stop taking it for the pain. If you find a doctor that will treat your pain you will be very lucky and very blessed.

Now I just live with the pain, like his leg will be squeezed and the cause, indicating that it doesn't always seem to work.
Quality of Question-Answer Thread

By nikgonz | Jan 18, 2008

20 Comments

AntiqueLady001 | Jun 12, 2013
To: TerrySa

Have you tried Lyrica or Gabapentin? They are good meds to take for the pain of nerve damage. I'm type 2 diabetic and after a triple bypass I developed diabetic neuropathy (nerve damage on the soles of my feet, then shooting pains in both my feet & legs. Gabapentin helps somewhat but the dosage needs to be increased.

Snowbird2002 | Jun 30, 2014
To: AntiqueLady001
Quality of Question-Answer Thread
Quality of Question-Answer Thread
Quality of Question-Answer Thread

Truth Discovery
Challenge (1): Noisy Input

- Raw textual data, unstructured
- Error introduced by extractor
Challenge (2): Long-tail Phenomenon

- Long-tail on both object and source sides
- Most questions have few answers
Challenge (3): Multiple Linked Truths

• Truths can be multiple, and they are correlated with each other
Challenge (4): Efficiency Issue

• Truth Discovery
  • iterative procedure

One Chinese Medical Q&A forum:
  • millions of registered patients
  • hundreds of thousands of doctors
  • thousands of new questions per day

• Medical QA
  • large-scale data
Overview of Our System

raw Data → Filtering → <Q, A> pairs → Entity Extraction → symptoms, diseases, drugs, etc → application

Extracted Knowledge → 

Entity Extraction

symptoms, diseases, drugs, etc
Q&A System

25-year-old, cough, fever

Extracted Knowledge by Our System

Bronchitis is 25%
Cold 45%
Tuberculosis 30%
Q&A System

25-year-old, cough, fever

Extracted Knowledge by Our System

Azithromycin
Albuterol
Rifampin

Bronchitis: 25%
Tuberculosis: 30%
Cold: 45%
Overview

1. Introduction

2. Truth Discovery: Veracity Analysis from Sources and Claims

3. Truth Discovery Scenarios

4. Veracity Analysis from Features of Sources and Claims

5. Applications

6. Open Questions and Resources

7. References
Open Questions

• Data with complex types and structures
• Theoretical analysis
• Efficiency of veracity analysis
• Interpretation and evaluation
• Application-specific challenges
Available Resources

• Survey for truth discovery
  • [Gupta&Han, 2011]
  • [Li et al., VLDB’12]
  • [Waguih et al., 2014]
  • [Waguih et al., ICDE’15]
  • [Li et al., 2016]

• Survey for source trustworthiness analysis
  • [Tang&Liu, WWW’14]
Available Resources

• Truth discovery data and code
  • http://lunadong.com/fusionDataSets.htm
  • http://cogcomp.cs.illinois.edu/page/resource_view/16
  • http://www.cse.buffalo.edu/~jing/software.htm
These slides are available at http://www.cse.buffalo.edu/~jing/talks.htm

KDD’16 Tutorial
Enabling the Discovery of Reliable Information from Passively and Actively Crowdsourced Data
- Budget allocation
- Privacy preservation
- Crowd sensing
-........
References


[Zhao&Han, QDB’12] B. Zhao, and J. Han. A probabilistic model for estimating real-valued truth from conflicting sources. In *Proc. of the VLDB workshop on Quality in Databases (QDB’12)*, 2012.


