Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources

Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, Wei Zhang

Google Inc.
@VLDB’2015
Motivation for Knowledge-Based Trust (KBT)

- Providing a new perspective to evaluate Web source quality
- What we have now--*Exogenous signals*
  - Link-based
  - Search log and click-through rate
  - Web spam
- Key idea: Evaluate trustworthiness of sources by the correctness of its factual information--*Endogenous signals*
Correctness of Factual Information

| Fact 1 | ✓ |
| Fact 2 | ✓ |
| Fact 3 | X |
| Fact 4 | ✓ |
| Fact 5 | X |
| Fact 6 | ✓ |
| Fact 7 | ✓ |
| Fact 8 | ✓ |
| Fact 9 | ✓ |
| Fact 10 | X |

Accu 0.7
How Can Trustworthiness Help?
Knowledge-Based Trust (KBT)

Trustworthiness in [0,1] for 5.6M websites and 119M webpages
Knowledge-Based Trust vs. PageRank

Correlated scores

Often tail sources w. high trustworthiness
I. Tale Sources w. Low PageRank May Provide Valuable Info

Among 100 sampled websites, 85 are indeed trustworthy.
Knowledge-Based Trust vs. PageRank

Often sources with high trustworthiness are correlated with high PageRank scores. Often sources with low accuracy are found in the lower left quadrant of the graph, indicating lower KBT (Knowledge-based Trust) scores.
## II. Popular Websites May Not Be Trustworthy

### Gossip Websites

*http://www.ebizmba.com/articles/gossip-websites*

<table>
<thead>
<tr>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.eonline.com">www.eonline.com</a></td>
</tr>
<tr>
<td>perezhilton.com</td>
</tr>
<tr>
<td>radaronline.com</td>
</tr>
<tr>
<td><a href="http://www.zimbio.com">www.zimbio.com</a></td>
</tr>
<tr>
<td>mediatakeout.com</td>
</tr>
<tr>
<td>gawker.com</td>
</tr>
<tr>
<td><a href="http://www.popsugar.com">www.popsugar.com</a></td>
</tr>
<tr>
<td><a href="http://www.people.com">www.people.com</a></td>
</tr>
<tr>
<td><a href="http://www.tmz.com">www.tmz.com</a></td>
</tr>
<tr>
<td><a href="http://www.fishwrapper.com">www.fishwrapper.com</a></td>
</tr>
<tr>
<td>celebrity.yahoo.com</td>
</tr>
<tr>
<td>wonderwall.msn.com</td>
</tr>
<tr>
<td>hollywoodlife.com</td>
</tr>
<tr>
<td><a href="http://www.wetpaint.com">www.wetpaint.com</a></td>
</tr>
</tbody>
</table>

14 out of 15 have a PageRank among top 15% of the websites

All have knowledge-based trust in bottom 50%
II. Popular Websites May Not Be Trustworthy
### III. Website Recommendation by Vertical

<table>
<thead>
<tr>
<th>Site</th>
<th>Accuracy</th>
<th># Triples</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>en.wikipedia.org</td>
<td>0.76</td>
<td>1,184</td>
<td>5.36</td>
</tr>
<tr>
<td><a href="http://www.cheesewiki.com">www.cheesewiki.com</a></td>
<td>0.84</td>
<td>230</td>
<td>4.55</td>
</tr>
<tr>
<td><a href="http://www.ranker.com">www.ranker.com</a></td>
<td>0.93</td>
<td>105</td>
<td>4.34</td>
</tr>
<tr>
<td>scratchpad.wikia.com</td>
<td>0.92</td>
<td>101</td>
<td>4.26</td>
</tr>
<tr>
<td><a href="http://www.marcellathecheesemonger.com">www.marcellathecheesemonger.com</a></td>
<td>0.90</td>
<td>109</td>
<td>4.25</td>
</tr>
<tr>
<td><a href="http://www.cheeseplatesf.com">www.cheeseplatesf.com</a></td>
<td>0.90</td>
<td>114</td>
<td>4.25</td>
</tr>
<tr>
<td>cheesetique.com</td>
<td>0.93</td>
<td>83</td>
<td>4.12</td>
</tr>
<tr>
<td><a href="http://www.lafromagerie.co.uk">www.lafromagerie.co.uk</a></td>
<td>0.93</td>
<td>62</td>
<td>3.87</td>
</tr>
<tr>
<td>adrian1974fulga.wordpress.com</td>
<td>0.90</td>
<td>71</td>
<td>3.83</td>
</tr>
<tr>
<td>epicurefoodscorp.com</td>
<td>0.78</td>
<td>131</td>
<td>3.81</td>
</tr>
<tr>
<td><a href="http://www.gourmetfoodstore.com">www.gourmetfoodstore.com</a></td>
<td>0.83</td>
<td>91</td>
<td>3.77</td>
</tr>
<tr>
<td><a href="http://www.sheridanscheesemongers.com">www.sheridanscheesemongers.com</a></td>
<td>0.85</td>
<td>71</td>
<td>3.65</td>
</tr>
<tr>
<td><a href="http://www.sfgate.com">www.sfgate.com</a></td>
<td>0.89</td>
<td>54</td>
<td>3.56</td>
</tr>
<tr>
<td>cheeseandchampagne.com</td>
<td>0.90</td>
<td>48</td>
<td>3.50</td>
</tr>
<tr>
<td><a href="http://www.sciencedirect.com">www.sciencedirect.com</a></td>
<td>0.69</td>
<td>161</td>
<td>3.50</td>
</tr>
<tr>
<td>about-france.com</td>
<td>0.98</td>
<td>34</td>
<td>3.48</td>
</tr>
<tr>
<td><a href="http://www.cookipedia.co.uk">www.cookipedia.co.uk</a></td>
<td>0.70</td>
<td>132</td>
<td>3.44</td>
</tr>
<tr>
<td><a href="http://www.cheese.com">www.cheese.com</a></td>
<td>0.86</td>
<td>50</td>
<td>3.39</td>
</tr>
<tr>
<td><a href="http://www.pennmac.com">www.pennmac.com</a></td>
<td>0.76</td>
<td>87</td>
<td>3.38</td>
</tr>
</tbody>
</table>
## III. Website Recommendation by Vertical

<table>
<thead>
<tr>
<th>Site</th>
<th>Accuracy</th>
<th># Triples</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ru.wikipedia.org</td>
<td>0.85</td>
<td>3,597</td>
<td>6.98</td>
</tr>
<tr>
<td>ja.wikipedia.org</td>
<td>0.73</td>
<td>3,032</td>
<td>5.88</td>
</tr>
<tr>
<td>en.wikipedia.org</td>
<td>0.53</td>
<td>36,489</td>
<td>5.57</td>
</tr>
<tr>
<td><a href="http://www.opensource/softwaredirectory.com">www.opensource/softwaredirectory.com</a></td>
<td>0.91</td>
<td>323</td>
<td>5.26</td>
</tr>
<tr>
<td>wpedia.goo.ne.jp</td>
<td>0.63</td>
<td>2,846</td>
<td>5.03</td>
</tr>
<tr>
<td><a href="http://www.ranker.com">www.ranker.com</a></td>
<td>0.81</td>
<td>308</td>
<td>4.66</td>
</tr>
<tr>
<td>packages.gentoo.org</td>
<td>0.65</td>
<td>942</td>
<td>4.45</td>
</tr>
<tr>
<td>uk.wikipedia.org</td>
<td>0.73</td>
<td>420</td>
<td>4.43</td>
</tr>
<tr>
<td>freecode.com</td>
<td>0.57</td>
<td>2,085</td>
<td>4.36</td>
</tr>
<tr>
<td><a href="http://www.file.net">www.file.net</a></td>
<td>0.65</td>
<td>756</td>
<td>4.34</td>
</tr>
<tr>
<td>gpo.zugaina.org</td>
<td>0.65</td>
<td>711</td>
<td>4.28</td>
</tr>
<tr>
<td>gentoobrowse.randomdan.homeip.net</td>
<td>0.80</td>
<td>208</td>
<td>4.27</td>
</tr>
<tr>
<td>whatis.techtarget.com</td>
<td>0.56</td>
<td>1,417</td>
<td>4.08</td>
</tr>
<tr>
<td><a href="http://www.starringthecomputer.com">www.starringthecomputer.com</a></td>
<td>0.87</td>
<td>101</td>
<td>4.02</td>
</tr>
<tr>
<td><a href="http://www.linuxlinks.com">www.linuxlinks.com</a></td>
<td>0.67</td>
<td>411</td>
<td>4.02</td>
</tr>
<tr>
<td>companies.findthebest.com</td>
<td>0.81</td>
<td>135</td>
<td>3.95</td>
</tr>
<tr>
<td><a href="http://www.fileinfo.com">www.fileinfo.com</a></td>
<td>0.53</td>
<td>1,770</td>
<td>3.95</td>
</tr>
<tr>
<td>file.downloadatoz.com</td>
<td>0.62</td>
<td>571</td>
<td>3.92</td>
</tr>
<tr>
<td><a href="http://www.zwodnik.com">www.zwodnik.com</a></td>
<td>0.55</td>
<td>1,187</td>
<td>3.92</td>
</tr>
<tr>
<td><a href="http://www.system-tray-cleaner.com">www.system-tray-cleaner.com</a></td>
<td>0.70</td>
<td>273</td>
<td>3.92</td>
</tr>
<tr>
<td>bitnami.com</td>
<td>0.70</td>
<td>266</td>
<td>3.91</td>
</tr>
</tbody>
</table>
Now, How to Compute KBT?
Key Idea in KBT

United States

From Wikipedia, the free encyclopedia
(Redirected from USA)

For other uses, see US (disambiguation), USA (disambiguation), and United States (disambiguation).

The United States of America (USA), commonly referred to as the United States (US), America or simply the States, is a federal republic consisting of 50 states and a federal district. The 48 contiguous states and the federal district of Washington, D.C., are in central North America between Canada and Mexico. The state of Alaska is the northwestern part of North America and the state of Hawaii is an archipelago in the mid-Pacific. The country also has five populated and nine unpopulated territories in the Pacific and the Caribbean. At 3.79 million square miles (9.83 million km²) in total and with around 317
d

Accu 0.7
Knowledge Vault—Probabilistic Knowledge Fusion

- #Triples: 3.0B (0.3B w. pr>=0.7)
- #URLs: 2.5B (28M Websites)
- #Extractors: 16

[SIGKDD, 2014]
[VLDB, 2014]
United States

From Wikipedia, the free encyclopedia
(Redirected from USA)

For other uses, see US (disambiguation), USA (disambiguation), and United States (disambiguation).

The United States of America (USA), commonly referred to as the United States (US), America or simply the States, is a federal republic consisting of 50 states and a federal district. The 48 contiguous states and the federal district of Washington, D.C., are in central North America between Canada and Mexico. The state of Alaska is the northwestern part of North America and the state of Hawaii is an archipelago in the mid-Pacific. The country also has five populated and nine unpopulated territories in the Pacific and the Caribbean. At 3.79 million square miles (9.83 million km²) in total and with around 317

Fact 1 ✓
Fact 2 ✓
Fact 3 ✗
Fact 4 ✓
Fact 5 ✗
Fact 6 ✓
Fact 7 ✓
Fact 8 ✓
Fact 9 ✓
Fact 10 ✗

Accu 0.7
KV Makes This Possible

United States

From Wikipedia, the free encyclopedia
(Redirected from USA)

For other uses, see US (disambiguation), USA (disambiguation), and United States (disambiguation).

The United States of America (USA), commonly referred to as the United States (US), America or simply the States, is a federal republic consisting of 50 states and a federal district. The 48 contiguous states and the federal district of Washington, D.C., are in central North America between Canada and Mexico. The state of Alaska is the northwestern part of North America and the state of Hawaii is an archipelago in the mid-Pacific. The country also has five populated and nine unpopulated territories in the Pacific and the Caribbean. At 3.79 million square miles (9.83 million km²) in total and with around 317

| Triple 1 | 1.0 |
| Triple 2 | 0.9 |
| Triple 3 | 0.3 |
| Triple 4 | 0.8 |
| Triple 5 | 0.4 |
| Triple 6 | 0.8 |
| Triple 7 | 0.9 |
| Triple 8 | 1.0 |
| Triple 9 | 0.7 |
| Triple 10 | 0.2 |
| ... | ... |

Accu 0.7
### Challenges

How to decide if a triple is indeed claimed by the source instead of an extraction error?
Extractions Can Be Wrong

- (Obama, nationality, Kenya)
  - 2087 extractions:
    - Example of a correct extraction
      - 2006: Obama In Kenya: I Am So Proud To Come Back Home - [VIDEO HERE.]
      - 2007: Michelle Obama Declares Obama Is Kenyan And America Is Mean - [VIDEO HERE.]
      - 2008: Michelle Obama Declares Barack Obama’s Home Country Is Kenya - [VIDEO HERE.]
      - FLASHBACK: Obama Is The Original Birther! Obama In 1991 Stated In His Own Bio He Was Born In Kenya. [DETAILS HERE.]
    - Example of a wrong extraction
      - [Link](http://www.monitor.co.ug/News/National/US+will+respect+winner+of+Kenya+election++Obama+says/-/688334/1685814/-/ksxagx/-/index.html)
      - US will respect winner of Kenya election, Obama says
Extractions Can Be Wrong

- (Obama, nationality, USA)

2481 extractions:

- Example of a correct extraction
  

- Example of a wrong extraction
  
KBT Strategies

1. Graphical model--predict at the same time
   a. extraction correctness
   b. triple correctness
   c. source accuracy
   d. extractor precision/recall

2. Un(Semi-)supervised learning (Bayesian)
   a. leverage source/extractor agreements
   b. trust a source/extractor w. high quality

3. Source/extractor hierarchy
   a. Break down “large” sources
   b. Group “small” sources
Graphical Model

Observations
- \( X_{ewdv} \): whether extractor e extracts from source w the \((d,v)\) item-value pair

Latent variables
- \( C_{wdv} \): whether source w indeed provides \((d,v)\) pair
- \( V_d \): the correct value(s) for d

Parameters
- \( A_w \): Trust of source w
- \( P_e \): Precision of extractor e
- \( R_e \): Recall of extractor e
Algorithm

Compute $Pr(W \text{ provides } T \mid \text{Extractor quality})$ by Bayesian analysis

Compute $Pr(T \mid \text{Source quality})$ by Bayesian analysis

Compute source accuracy

Compute extractor precision and recall

E-Step

M-Step
United States

From Wikipedia, the free encyclopedia
(Redirected from USA)

For other uses, see US (disambiguation), USA (disambiguation), and United States (disambiguation).

The United States of America (USA), commonly referred to as the United States (US), America or simply the States, is a federal republic consisting of 50 states and a federal district. The 48 contiguous states and the federal district of Washington, D.C., are in central North America between Canada and Mexico. The state of Alaska is the northwestern part of North America and the state of Hawaii is an archipelago in the mid-Pacific. The country also has five populated and nine unpopulated territories in the Pacific and the Caribbean. At 3.79 million square miles (9.83 million km²) in total and with around 317
Predicting Extraction and Triple Correctness

- (Obama, nationality, Kenya)
  
  2087 extractions:
  
  - Example of a correct extraction \( \text{Pr}_{\text{extCorr}}=0.792 \)
    
    
    2007: Michelle Obama Declares Obama Is Kenyan And America Is Mean - VIDEO HERE.
    2008: Michelle Obama Declares Barack Obama’s Home Country Is Kenya - VIDEO HERE.
    FLASHBACK: Obama Is The Original Birther! Obama In 1991 Stated In His Own Bio He Was Born In Kenya. DETAILS HERE.

  - Example of a wrong extraction \( \text{Pr}_{\text{extCorr}}=0.130 \)
    
    http://www.monitor.co.ug/News/National/US+will+respect+winner+of+Kenya+election++Obama+says/-/688334/1685814/-/ksxagx/-/index.html
    
    US will respect winner of Kenya election, Obama says

- \( \text{Pr}_{\text{tripleCorr}}=0 \) (not enough support)
Predicting Extraction and Triple Correctness

- (Obama, nationality, USA)
  2481 extractions:
  - Example of a correct extraction \((\text{Pr}_{\text{extCorr}}=0.999)\)
    - [Link](http://www.dogonews.com/2009/10/9/a-nobel-prize-for-our-awesome-president)
  - Example of a wrong extraction \((\text{Pr}_{\text{extCorr}}=0.261)\)

- \(\text{Pr}_{\text{tripleCorr}}=1\) (higher support)
Predicting Extraction and Triple Correctness

Distribution of providers for Kenya and USA

Distribution of Pr_extCorr for Kenya vs USA
Predicting Extraction and Triple Correctness

Triple Correctness Prediction

Extraction Correctness Prediction

- Prediction
- Ideal

- Type error
- Freebase triple
Predicting Triple Correctness

PR-Curves

ROC-Curves
What is the Future of KBT?
Future Works

1. Extraction is still very sparse
   a. 74% URLs each contributes fewer than 5 triples
   b. We compute reliable KBT for <20% websites and <<5% webpages

2. Extraction is of low quality
   a. Overall accuracy is as low as 11.5%
   b. Low accuracy for some good sources because of undetected extraction errors

Call to arms --
Leave NO Valuable Data Behind
Press Coverage of the Paper

Press Coverage

Hooray, Truths on the Way!
... I read with interest your recent paper on KBT ... Actually, that’s false — I tried to read it, and did read all of the parts that weren't numbers and Greek characters. It is quite an interesting proposal, though.

I’m writing because XXX published a piece claiming that YYY would be injured under a ranking system that took KBT into account. They must have got that from footnote 16 in your paper ...

I’m writing with a simple request: Can you provide me with the XXX’s KBT score and percentile ranking, and how it compares to YYY’s?
Google has developed a technology to tell whether ‘facts’ on the Internet are true

It’s unclear exactly what Google plans to do with this new technology, if anything at all. Still, even the possibility of a search engine that evaluates truth is a pretty incredible breakthrough. And it definitely gives new meaning to the phrase “let me Google that for you.”
THANK YOU!