Knowledge Vault: a web-scale approach to probabilistic knowledge fusion

Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, Wei Zhang

Google (Machine Intelligence group)
Outline of the talk

1. Knowledge Graph
2. Knowledge Vault
3. Fact mining from the web
4. Fact mining from graphs
5. Knowledge Fusion
A Knowledge Graph is a multi-graph where nodes = entities, edges = relations

- Kobe Bryant
- NY Knicks
- LA Lakers
- Pau Gasol
- playFor
- teamInLeague
- opponent
- teammate
- playInLeague
Example Knowledge Graphs

- Google’s KG
- Walmart’s Kosmix
- Microsoft’s Satori
- Facebook’s Entity Graph
- Freebase
- Yago
- NELL: Never-Ending Language Learning
- DBpedia
Freebase is created by fusing structured data sources and human contributions.
The long tail of knowledge

- FB is very large (40M nodes, 637M edges)
- But it still very incomplete:
  - We are missing many edges (facts)
  - We are also missing many nodes (entities)
  - We are also missing many edge types (schema)

<table>
<thead>
<tr>
<th>Relation</th>
<th>% unknown in Freebase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profession</td>
<td>68%</td>
</tr>
<tr>
<td>Place of birth</td>
<td>71%</td>
</tr>
<tr>
<td>Nationality</td>
<td>75%</td>
</tr>
<tr>
<td>Education</td>
<td>91%</td>
</tr>
<tr>
<td>Spouse</td>
<td>92%</td>
</tr>
<tr>
<td>Parents</td>
<td>94%</td>
</tr>
</tbody>
</table>

This talk
Outline of the talk

1. Knowledge Graph
2. Knowledge Vault
3. Fact mining from the web
4. Fact mining from graphs
5. Knowledge Fusion
From Knowledge Graph to Knowledge Vault

- There are many groups at Google working on enlarging KG while maintaining high precision.
- KV is an exploratory research project to investigate other points along the precision-recall curve.
- KV automatically extracts facts from public web sources.
- KV embraces the inherent uncertainty associated with this process (every fact has associated confidence and provenance info).
Previous projects on automatically building KBs (eg NELL, YAGO) predict facts based on text

Pr(<s, r, o>=1|D)

“Kobe Bryant, "Kobe "Kobe Bryant
the franchise player of once again saved man of the match for

the Lakers”
his team”
Los Angeles”
KV: Predict new facts based on text AND existing edges in FB

Pr(<s, r, o>=1|D)

“Kobe Bryant, “Kobe “Kobe Bryant
the franchise player of once again saved man of the match for

the Lakers” his team” Los Angeles”
KV is 50x bigger than comparable KBs

Total # facts in KV > 2.5B

<table>
<thead>
<tr>
<th>Name</th>
<th># Entity types</th>
<th># Entity instances</th>
<th># Relation types</th>
<th># Confident facts (relation instances)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge Vault (KV)</td>
<td>1100</td>
<td>45M</td>
<td>4469</td>
<td>302M</td>
</tr>
<tr>
<td>DeepDive [32]</td>
<td>4</td>
<td>2.7M</td>
<td>34</td>
<td>7M&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>NELL [8]</td>
<td>271</td>
<td>5.19M</td>
<td>306</td>
<td>0.435M&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>PROSPERA [30]</td>
<td>11</td>
<td>N/A</td>
<td>14</td>
<td>0.1M</td>
</tr>
<tr>
<td>YAGO2 [19]</td>
<td>350,000</td>
<td>9.8M</td>
<td>100</td>
<td>4M&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Freebase [4]</td>
<td>1,500</td>
<td>40M</td>
<td>35,000</td>
<td>637M&lt;sup&gt;d&lt;/sup&gt;</td>
</tr>
<tr>
<td>Knowledge Graph (KG)</td>
<td>1,500</td>
<td>570M</td>
<td>35,000</td>
<td>18,000M&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Open IE (e.g., Mausam et al., 2012)
5B assertions (Mausam, Michael Schmitz, personal communication, October 2013)

302M with Prob > 0.9
381M with Prob > 0.7
Uses for KV's uncertain triples

probably false triples removed from KG

possibly false triples used for error analysis

possibly true triples used as weak signals

probably true triples uploaded to KG
Outline of the talk

1. Knowledge Graph
2. Knowledge Vault
3. Fact mining from the web
4. Fact mining from graphs
5. Knowledge Fusion
Fact extraction from the web

Webmaster annotations → Extractors → Fusion
NL text → Extractors
Page structure → Extractors
Tables → Extractors

Knowledge Vault
Fact extraction from text (TXT)

- First identify named entities (entity linkage).
- Then classify verb phrase as one of 2000 relations

Patrick Newport, who has been working at IHS Global Insight, noted...

The result is a probabilistic triple:

$$\text{Pr}(\text{<subject, reln, object>}=1 \mid \text{text})$$

Classifier trained using distant supervision.*

Details: see eg tutorial by Ralph Grishman (NYU): “Information Extraction: Capabilities and Challenges”, 2012

* Mintz et al, RANLP 2009
Fact extraction from DOM trees*

- First identify named entities on page
- Then classify X-path connecting each entity pair as one of 2000 relations

* Cafarella et al, CACM’11
Fact extraction from tables (TBL)*

<table>
<thead>
<tr>
<th>Year</th>
<th>Illustrator</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>Mary Azarian</td>
<td><em>Snowflake Bentley</em></td>
</tr>
<tr>
<td>1954</td>
<td>Ludwig Bemelmans</td>
<td><em>Madeline’s Rescue</em></td>
</tr>
<tr>
<td>1983</td>
<td>Marcia Brown</td>
<td><em>Shadow</em></td>
</tr>
<tr>
<td>1962</td>
<td>Marcia Brown</td>
<td><em>Once a Mouse</em></td>
</tr>
<tr>
<td>1955</td>
<td>Marcia Brown</td>
<td><em>Cinderella, or the Little Glass Slipper</em></td>
</tr>
<tr>
<td>1943</td>
<td>Virginia Lee Burton</td>
<td><em>The Little House</em></td>
</tr>
<tr>
<td>1980</td>
<td>Barbara Cooney</td>
<td><em>Ox-Cart Man</em></td>
</tr>
<tr>
<td>1959</td>
<td>Barbara Cooney</td>
<td><em>Chanticleer and the Fox</em></td>
</tr>
</tbody>
</table>

Squares are CVT nodes

* Cafarella et al, VLDB’08
Fact extraction from schema.org annotation (ANO)

- About 20% of webpages have machine-readable annotations of commercial events, products, etc.
- Automatically map to KG schema.
- We still need to do entity linking.
Combine outputs from all extractors

- Train binary classifier on $f(t) = [\text{score-txt}(t), \#\text{txt}(t), \ldots]$ using distant supervision.
- Platt scaling to get calibrated probabilities.

Extractors

- Webmaster annotations
- Tables
- NL text
- Page structure
ROC for each extraction system

- ANO (0.73)
- TBL (0.76)
- TXT (0.85)
- DOM (0.87)
- FUSED-EX (0.92)
Confidence of true facts rises given more evidence
Outline of the talk

1. Knowledge Graph
2. Knowledge Vault
3. Fact mining from the web
4. Fact mining from graphs
5. Knowledge Fusion
Mining facts from graphs

Web

Extractors

Fusion

Priors
Link prediction using tensor factorization

- Many methods have been used to fill in missing values in binary matrices, e.g., tensor factorization associates a low-dimensional vector with every row and column.
(Deep) neural network for link prediction

- Represent each entity and relation by its own low-dimensional (100D) embedding vector.
- Stack together, feed into neural net.
- Train model to maximize log-likelihood of observed positive and negative triples.
- Outperforms neural tensor model (Socher et al).
Path Ranking Algorithm [Lao et al., EMNLP11]

CityLocatedInCountry(Pittsburgh) = ?

CityLocatedInCountry, Path Ranking Algorithm [Lao et al., EMNLP11]

Feature = Typed Path

<table>
<thead>
<tr>
<th>Feature Value</th>
<th>Logistic Regression Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.32</td>
</tr>
<tr>
<td>0.6</td>
<td>0.20</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

CityLocatedInCountry(Pittsburgh) = U.S.  p=0.58

Figure courtesy of Tom Mitchell and Partha Talukdar
Example of paths / rules learned by PRA

CityLocatedInCountry\( (city, country) \):

7 of the 2985 learned paths

8.04 cityliesonriver, cityliesonriver\(^{-1}\), citylocatedincountry

5.42 hasofficeincity\(^{-1}\), hasofficeincity, citylocatedincountry

4.98 cityalsoknownas, cityalsoknownas, citylocatedincountry

2.85 citycapitalofcountry, citylocatedincountry\(^{-1}\), citylocatedincountry

2.29 agentactsinlocation\(^{-1}\), agentactsinlocation, citylocatedincountry

1.22 statehascapital\(^{-1}\), statelocatedincountry

0.66 citycapitalofcountry
PRA similar in performance to neural network
Outline of the talk

1. Knowledge Graph
2. Knowledge Vault
3. Fact mining from the web
4. Fact mining from graphs
5. Knowledge Fusion
Fusing web extractions with graph priors
Example: (Barry Richter, studiedAt, UW-Madison)

“In the fall of 1989, Richter accepted a scholarship to the University of Wisconsin, where he played for four years and earned numerous individual accolades ...”

“The Polar Caps' cause has been helped by the impact of knowledgeable coaches such as Andringa, Byce and former UW teammates Chris Tancill and Barry Richter.”

➤ Web extraction confidence: 0.14

Freebase

<Barry Richter, born in, Madison>
<Barry Richter, lived in, Madison>

➤ Final belief (fused with prior): 0.61
Summary and future work

• KV has 2.5B triples automatically extracted from the web.
• Combining web mining and graph mining can improve precision.
• Work in progress
  ▪ Discovering new entities
    • Clustering open IE extractions, CIKM 2014
    • Robust wrapper induction for long-tail verticals (work in progress)
  ▪ Discovering new relations
    • Clustering open IE extractions, CIKM 2014
    • “Biperpedia”, VLDB 2014
  ▪ Assessing trust-worthiness of web sites: VLDB 2014
  ▪ Common sense fact mining eg “apples” (work in progress)
Application 1: Knowledge Panels

Augmenting the presentation with relevant facts
Application 2: Related Entities

Gray Line New York Sightseeing Tours, Cruises & Attractions
www.newyorksightseeing.com
Gray Line New York
New York’s famous Empire State Building, a New York City and a National Historic to Top of the Rock 1-hour Statue of Liberty New York Harbor Cruise!
Nyc double decker tours - Loops Tour Map - All Loops Tour Plus - Contact Us

New York attractions: The 50 best sights and attractions in ...
www.timeout.com/newyork/attractions.../new-york-attractions - Top Out
by Amy Ryan - Apr 25, 2013 - Sights like the Empire State Building and the Statue of Liberty are perennial favorites, but we’ve also highlighted newcomers and lesser-known

BLDG 92 - Free attractions in New York - Brooklyn Flea - Chrysler Building

New York City Tours and Attractions - NYC Sightseeing ...

New York: Sightseeing in NYC - TripAdvisor
www.tripadvisor.com - New York (NY) - Before You Go - TripAdvisor
Inside New York: Sightseeing in NYC - Before you visit New York, visit ... the main sights of the New York City harbor, including the Statue of Liberty, Ellis Island ...

NYC Sightseeing Tour | New York City Double Decker Tour...
skylinenewyork.com
See NYC’s landmarks your way with our Hop-on. ... The most historic neighborhood in New York City, downtown Manhattan features the Empire State Building ...

City Sightseeing New York, Hop On - Hop Off Bus Tours
www.citysightseeing.com/tours/united-states-of-new-york.htm
Choose your own way in which you use your ticket according to your own itinerary.
There are three tour routes to choose from, allowing you to explore.

Top 25 New York City Tours - New York Magazine
Aug 16, 2013 - Views From the Top: The Big Apple Tour. See the city streets from a
Application 3: Structured Graph Search

Figure courtesy of Antoines Bordes (Facebook)
Application 4: Factoid Question Answering

Google

EVI
(Amazon)

Siri
(Apple)
The yield from different extraction systems
Overlap between extractors

- **TXT (301M)**
  - 1.1M connections to **TBL (10M)**
  - 0.3M connections to **DOM (1280M)**

- **DOM (1280M)**
  - 1.5M connections to **TBL (10M)**
  - 1.7M connections to **ANO (145M)**

- **TBL (10M)**
  - 13K connections to **ANO (145M)**

- **ANO (145M)**