Challenges and Innovations in Building a Product Knowledge Graph

XIN LUNA DONG, AMAZON

JANUARY, 2018
Product Graph vs. Knowledge Graph
Knowledge Graph Example for 2 Movies

- **“Forrest Gump”**
  - Directed by: mid127
  - Starring: mid128, mid129
- **“Larry Crowne”**
  - Directed by: mid127
  - Starring: mid128, mid129

- **People**
  - Tom Hanks
  - Julia Roberts
  - Robin Wright
  - Robin Wright Penn
  - 罗宾·怀特

- **Birth Date**
  - July 9th, 1956
Knowledge Graph in Search

List of Tom Hanks performances - Wikipedia
https://en.wikipedia.org/wiki/List_of_Tom_Hanks_performances

Tom Hanks (@tomhanks) - Twitter
https://twitter.com/tomhanks

And don't miss this songstress at the famous Cafe Carlyle. Through Saturday nite! Hanx @RitaWilson pic.twitter.com/J7OXIbf.. 12 hours ago · Twitter
Beware! Cram self-serving Social Media Post! This book goes on sale tomorrow! Hanx pic.twitter.com/VZEqPKL.. 16 hours ago · Twitter
Lost (g)love. Looking for a mate. Good luck. Hanx. pic.twitter.com/ApH7tEG.. 1 day ago · Twitter

Tom Hanks
American actor

Thomas Jeffrey Hanks is an American actor and filmmaker. He is known for his various comedic and dramatic film roles, including Splash, Big, Turner & Hooch, A League of Their Own, Sleepless in Seattle, ... Wikipedia

Born: July 9, 1956 (age 61), Concord, CA
Awards: Academy Award for Best Actor, MORE
Spouse: Rita Wilson (m. 1988), Samantha Lewes (m. 1978–1987)
TV shows: Bosom Buddies, Celebrity Jeopardy, MORE
Knowledge Graph in Personal Assistant

List of officially released compilations and
[92][93][94][95]
- Portrait of Michael Jackson / Portrait of Jackson 5 (1973)
- Motown Superstar Series, Vol. 7 (1980)
- Superstar (1980)
- Michael Jackson & The Jackson 5 (1983)
- Ain't No Sunshine (1984)
- The Great Love Songs of Michael Jackson (1984)
- Ben / Got to Be There (1986)
- Looking Back to Yesterday (1986)
- The Original Soul of Michael Jackson (1987)
- Rockin' Robin (1993)
- Dangerous – The Remix Collection (1993)
- Michael Jackson Story (1996)
- Master Series (1997)
- Ghosts – Deluxe Collector Box Set (1997)
- Got to Be There / Forever, Michael (1999)
- Bad / Thriller (2000)
- Forever, Michael / Music & Me / Ben (2000)
- Classic – The Universal Masters Collection (2001)

Alexa, play the music by Michael Jackson
Product Graph

- Mission: To answer any question about products and related knowledge in the world
Product Graph vs. Knowledge Graph

(A) Generic KG

(B) Generic KG

(C) Generic KG

PG
Knowledge Graph Example for 2 Movies

Forrest Gump

Larry Crowne

“Tom Hanks”

“Julia Roberts”

July 9th, 1956

“Robin Wright Penn”

“Robin Wright”

“Julia Roberts”

“Tom Hanks”

July 9th, 1956
Product Graph vs. Knowledge Graph

mid567
mid568
mid569
mid570
mid571
mid345
mid346
mid128
mid127
mid129
B0035QUXWQ
B0035QUXWR
Digital Movie
DVD
Blu-ray
B0067XLIG8
B0067XLIG4
“Robin Wright”
“Robin Wright Penn”
“罗宾·怀特”
“Tom Hanks”
July 9th, 1956
“Julia Roberts”
Person

“Forrest Gump”
“Larry Crowne”

starring
starring
directed_by

name
name
name
name
name
name

ASIN
ASIN
ASIN
ASIN
ASIN

product
product
product
product

B0035QUXWQ
B0067XLIG4

Digital Movie
Blu-ray

B0035QUXWR
B0067XLIG8
Another Example of Product Graph
Knowledge Graph vs. Product Graph

(A) Generic KG

(B) Generic KG
- PG (Movie, Music, Book)

(C) Generic KG
- Movie, Music, Book, etc.
- Product Graph
But, Is The Problem Harder?

Generic KG

Movie, Music, Book, etc.

Product Graph
Challenges in Building Product Graph I

- No major sources to curate product knowledge from
  - Wikipedia does not help too much
  - A lot of structured data buried in text descriptions in Catalog
  - Retailers gaming with the system so noisy data
Challenges in Building Product Graph II

- Large number of new products everyday
  - Curation is impossible
  - Freshness is a big challenge
Challenges in Building Product Graph III

- Large number of product categories
  - A lot of work to manually define ontology
  - Hard to catch the trend of new product categories and properties
How to Build a Product Graph?
Where is Knowledge from?

Product Graph

Focus on Quality

amazon

WWW

Yes We’re OPEN

Structured Data
Which ML Model Works Best?
Which ML Model Works Best?

Tree-based models

Neural network
Research Philosophy

**Roofshots**: Deliver incrementally and make production impacts

**Moonshots**: Strive to apply and invent the state-of-the-art
I. Extracting Knowledge from Semi-Structured Data on the Web
I. Extracting Knowledge from Semi-Structured Data on the Web

Knowledge Vault @ Google showed big potential from DOM-tree extraction [Dong et al., KDD’14][Dong et al., VLDB’14]

<table>
<thead>
<tr>
<th></th>
<th>Accu</th>
<th>Accu (conf ≥ .7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TXT (301M)</td>
<td>0.36</td>
<td>0.52</td>
</tr>
<tr>
<td>DOM (1280M)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TBL (10M)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANO (145M)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Accu</th>
<th>Accu (conf ≥ .7)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.43</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>0.09</td>
<td>0.62</td>
</tr>
</tbody>
</table>
I. Extracting Knowledge from Web—Annotation-Based DOM Extraction

Extracted relationships

- (Top Gun, type.object.name, “Top Gun”)
- (Top Gun, film.film.genre, Action)
- (Top Gun, film.film.directed_by, Tony Scott)
- (Top Gun, film.film.starring, Tom Cruise)
- (Top Gun, film.film.runtime, “1h 50min”)
- (Top Gun, film.film.release_Date_s, “16 May 1986”)

Annotation-based knowledge extraction
I. Extracting Knowledge from Web—Annotation-Based DOM Extraction

Alexa, When did Padme Amidala die?
What model is R2D2?
Who is Luke Skywalker’s master?
Where is Boba Fett from?
Who is Darth Vader’s apprentice?
I. Extracting Knowledge from Web—Distantly Supervised DOM Extraction

Annotation-based knowledge extraction

Distantly supervised web extraction
I. Extracting Knowledge from Web—Distantly Supervised DOM Extraction

Entity Identification → Automatic Annotation → Training

Automatic Label Generation

Movie entity

Genre

Release Date

Runtime

Extracted triples

- (Top Gun, type.object.name, “Top Gun”)
- (Top Gun, film.film.genre, Action)
- (Top Gun, film.film.directed_by, Tony Scott)
- (Top Gun, film.film.starring, Tom Cruise)
- (Top Gun, film.film.runtime, “1h 50min”)
- (Top Gun, film.film.release_Date_s, “16 May 1986”)
I. Extracting Knowledge from Web—Distantly Supervised DOM Extraction

Extraction on IMDb

<table>
<thead>
<tr>
<th>Predicate</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type.object.name (“name”)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>People.person.place_of_birth</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Common.topic.alias</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Film.actor.film</td>
<td>0.98</td>
<td>0.47</td>
</tr>
<tr>
<td>Film.director.film</td>
<td>0.98</td>
<td>0.91</td>
</tr>
<tr>
<td>Film.producer.film</td>
<td>0.89</td>
<td>0.57</td>
</tr>
<tr>
<td>Film.writer.film</td>
<td>0.96</td>
<td>0.60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predicate</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV.tv_series_episode.episode_number</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TV.tv_series_episode.season_number</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Film.film.directed_by</td>
<td>0.99</td>
<td>1</td>
</tr>
<tr>
<td>Film.film.written_by</td>
<td>1</td>
<td>0.98</td>
</tr>
<tr>
<td>Film.film.genre</td>
<td>0.90*</td>
<td>1</td>
</tr>
<tr>
<td>Film.film.starring</td>
<td>1</td>
<td>0.97</td>
</tr>
<tr>
<td>TV.tv_series_episode.series</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

1. Very high extraction precision
2. Extracting triples with new entities

*Ground truth is incomplete. Manual inspection suggests close to 100% accuracy.
I. Extracting Knowledge from Web—Distantly Supervised DOM Extraction


<table>
<thead>
<tr>
<th>Site</th>
<th>Title</th>
<th>Director(s)</th>
<th>Genre(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>allmovies</td>
<td>1</td>
<td>1</td>
<td>0.71</td>
</tr>
<tr>
<td>amctv</td>
<td>1</td>
<td>0.98</td>
<td>0.95</td>
</tr>
<tr>
<td>boxofficemojo</td>
<td>1</td>
<td>1</td>
<td>0.97</td>
</tr>
<tr>
<td>hollywood</td>
<td>1</td>
<td>0.94</td>
<td>0.91</td>
</tr>
<tr>
<td>iheartmovies</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>IMDB</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>metacritic</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>MSN</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>rotten tomatoes</td>
<td>1</td>
<td>1</td>
<td>0.91</td>
</tr>
<tr>
<td>yahoo</td>
<td>1</td>
<td>1</td>
<td>0.94</td>
</tr>
</tbody>
</table>
I. Distantly Supervised DOM Extraction
Which ML Model Works Best?

- Logistic regression: Best results (20K features on one website)
- Random forest: lower precision and recall
I. Extracting Knowledge from Semi-Structured Data on the Web

Annotation-based knowledge extraction

Distantly supervised web extraction

OpenIE DOM extraction

Nearly-automatic interactive extraction on any new vertical

Annotation-based knowledge extraction
II. Extracting Knowledge from Product Profiles in Amazon Catalog

<table>
<thead>
<tr>
<th>name</th>
<th>form</th>
<th>scent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tide Detergent with Febreze Freshness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gain Apple Mango Tango Liquid Laundry Detergent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gain Joyful Expression Powder Detergent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tide PODS Original Scent HE Turbo Laundry Detergent Pacs 81-load Tub</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tide PODS Free &amp; Gentle HE Turbo Laundry Detergent Pacs 35-load Bag</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
II. Open Attribute Extraction by Named Entity Recognition

- **B**: Beginning of entity
- **I**: Inside of entity
- **O**: Outside of entity
- **E**: End of entity

$x = \{w_1, w_2, ..., w_n\}$ input sequence

$y = \{t_1, t_2, ..., t_n\}$ tagging decision

```
<table>
<thead>
<tr>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
<th>t6</th>
<th>t7</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>O</td>
<td>E</td>
<td>B</td>
<td>E</td>
<td>O</td>
<td>E</td>
<td></td>
</tr>
</tbody>
</table>
```

- beef
- meal
- &
- ranch
- raised
- lamb
- recipe

$X$
II. Open Attribute Extraction by NER — Which ML Model Works Best?

Recurrent Neural Network, CRF, Attention

- Hidden
  200 units

- Self Attention

- Embedding + BiLSTM
  glove embedding 50, 200 units

- Word Index
  5645 words

- ranch
- raised
- beef
- flavor
II. Open Attribute Extraction by NER — Adding Active Learning

Training
- 500 Sentences
- 7927 Words
- 944 Flavors

Testing
- 600 Sentences
- 7896 Words
- 786 Flavors

Different flavors from Training data
II. Open Attribute Extraction by NER — Attentions Help Find Contexts
II. Extracting Knowledge from Product Profiles in Amazon Catalog

- Product profile extraction
- Automatically building a shallow KG
- Open aspect extraction
- Review extraction & sentiment analysis
Take Aways

- We aim at building an authoritative knowledge graph for all products in the world
- We shoot for roofshot and moonshot goals to realize our mission
- There are many exciting research problems that we are tackling