Construction and Querying of Large-scale Knowledge Bases

Xiang Ren\textsuperscript{1}  Yu Su\textsuperscript{2}  Xifeng Yan\textsuperscript{2}

University of Southern California\textsuperscript{1}
University of California, Santa Barbara\textsuperscript{2}
Tutorial website:
http://xren7.web.engr.illinois.edu/tutorial-cikm17.html

Slides, code, datasets, references
Turning Unstructured Text Data into Structures

Unstructured Text Data (account for ~80% of all data in organizations)

Knowledge & Insights

(Chakraborty, 2016)
Reading the reviews: From Text to Structured Facts

This **hotel** is my favorite **Hilton property** in **NYC**! It is located right on 42nd street near **Times Square**, it is close to all subways, **Broadways shows**, and next to great restaurants like **Junior’s Cheesecake**, **Virgil’s BBQ** and many others.

--- **TripAdvisor**

**Structured Facts**

1. “Typed” entities
2. “Typed” relationships
Why Text to Structures?

Structured Search & Exploration

- City contains "ist" ×
- Category equals "Friends" ×
- Birthday on 09/04/2000 × Age = 30 ×
- Lastname equals "plugins" ×
- Is active □ Yes □ No □

Graph Mining & Network Analysis

Pattern / Association Rule Mining

Structured Feature Generation
A Product Use Case: Finding “Interesting Hotel Collections”

Technology Transfer to TripAdvisor

Features for “Catch a Show” collection
1. broadway shows
2. beacon theater
3. broadway dance center
4. broadway plays
5. david letterman show
6. radio city music hall
7. theatre shows

Features for “Near The High Line” collection
1. high line park
2. chelsea market
3. highline walkway
4. elevated park
5. meatpacking district
6. west side
7. old railway

Grouping hotels based on structured facts extracted from the review text

Prior Art: Extracting Structures with Repeated Human Effort

Extraction Rules
Machine-Learning Models

Structured Facts
- Broadways shows
- NYC
- Times square
- hotel
- Hilton property

Labeled data

Text Corpus

This hotel is my favorite Hilton property in NYC! It is located right on 42nd street near Times Square, it is close to all subways, Broadways shows, and next to many great ...

... We had a room facing *Times Square* and a room facing the *Empire State Building*, The location is close to everything and we love ...

Stanford CoreNLP
CMU NELL
UW KnowItAll
USC AMR
IBM Alchemy APIs
Google Knowledge Graph
Microsoft Satori
...
This Tutorial: Effort-Light StructMine

- Enables **quick** development of applications over various corpora
- Extracts **complex** structures without introducing human error
Effort–Light StructMine: Where Are We?

A Review of Previous Efforts

**Human labeling effort**

- **Hand-crafted Systems**
  - UCB Hearst Pattern, 1992
  - NYU Proteus, 1997

- **Supervised learning systems**
  - Stanford CoreNLP, 2005 - present
  - UT Austin Dependency Kernel, 2005
  - IBM Watson Language APIs

- **Weakly-supervised learning systems**
  - CMU NELL, 2009 - present
  - UW KnowItAll, Open IE, 2005 - present
  - Max-Planck YAGO, 2008 - present

- **Distantly-supervised Learning Systems**
  - Stanford DeepDive, MIML-RE 2012 - present
  - UW FIGER, MultiR, 2012

- **Effort-Light StructMine**
  - (WWW’15, KDD’15, KDD’16, EMNLP’16, WWW’17, ...)

**Feature engineering effort**
“Distant” Supervision: What Is It?

“Matchable” structures: entity names, entity types, typed relationships ...

Freely available!
- Common knowledge
- Life sciences
- Art ...

Rapidly growing!

Text corpus

Knowledge Bases

“Un-matchable”

(Mintz et al., 2009), (Riedek et al., 2010), (Lin et al., 2012), (Ling et al., 2012), (Surdeanu et al., 2012), (Xu et al., 2013), (Nagesh et al., 2014), ...

Learning with Distant Supervision: Challenges

1. Sparsity of “Matchable”
   - Incomplete knowledge bases
   - Low-confidence matching

2. Accuracy of “Expansion”
   - For “matchable”: Are all the labels assigned accurately?
   - For “un-matchable”: How to perform inference accurately?

(Ren et al., KDD’15)

It is my favorite city in the United States

The United States needs a new strategy to meet this challenge
**Effort-Light StructMine: Contributions**

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Solution Idea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparsity of “Matchable”</td>
<td>Effective expansion from “matchable” to “un-matchable”</td>
</tr>
<tr>
<td>Accuray of “Expansion”</td>
<td>Pick the “best” labels based on the context (for both “matchable” and “un-matchable”)</td>
</tr>
</tbody>
</table>

**Challenge Solution Idea**

**Harness the “data redundancy” using graph-based joint optimization**

- **Text corpus**
- **Wikipedia**
- **DBpedia**
- **Wikidata**
- **Freebase**

**Government Location**

- It is my favorite city in the United States.

- The United States needs a new strategy to meet this challenge.
Effort-Light StructMine: Methodology

Data-driven text segmentation (SIGMOD’15, WWW’16) → Entity names & context units

Structures from the remaining unlabeled data

Learning Corpus-specific Model (KDD’15, KDD’16, EMNLP’16, WWW’17) → Partially-labeled corpus

Knowledge bases
Knowledge Base Querying
Transformation in Information Search

Desktop search

Mobile search

“Which hotel has a roller coaster in Las Vegas?”

Lengthy Documents? Direct Answers!

Answer: New York-New York hotel

Surge of mobile Internet use in China

Source: China Internet Network Information Center
Application: Facebook Entity Graph

People, Places, and Things

Facebook’s knowledge graph (entity graph) stores as entities the users, places, pages and other objects within the Facebook.

Connecting

The connections between the entities indicate the type of relationship between them, such as friend, following, photo, check-in, etc.
Structured Query: RDF + SPARQL

**Triples in an RDF graph**

<table>
<thead>
<tr>
<th>Subject</th>
<th>Predicate</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barack_Obama</td>
<td>parentOf</td>
<td>Malia_Obama</td>
</tr>
<tr>
<td>Barack_Obama</td>
<td>parentOf</td>
<td>Natasha_Obama</td>
</tr>
<tr>
<td>Barack_Obama</td>
<td>spouse</td>
<td>Michelle_Obama</td>
</tr>
<tr>
<td>Barack_Obama_Sr.</td>
<td>parentOf</td>
<td>Barack_Obama</td>
</tr>
</tbody>
</table>

**SPARQL query**

```
SELECT ?x WHERE {
  Barack_Obama_Sr. parentOf ?y .
  ?y parentOf ?x .
}
```

**Answer**

```
<Malia_Obama>
<Natasha_Obama>
```
Why Structured Query Falls Short?

<table>
<thead>
<tr>
<th>Knowledge Base</th>
<th># Entities</th>
<th># Triples</th>
<th># Classes</th>
<th># Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freebase</td>
<td>45M</td>
<td>3B</td>
<td>53K</td>
<td>35K</td>
</tr>
<tr>
<td>DBpedia</td>
<td>6.6M</td>
<td>13B</td>
<td>760</td>
<td>2.8K</td>
</tr>
<tr>
<td>Google Knowledge Graph*</td>
<td>570M</td>
<td>18B</td>
<td>1.5K</td>
<td>35K</td>
</tr>
<tr>
<td>YAGO</td>
<td>10M</td>
<td>120M</td>
<td>350K</td>
<td>100</td>
</tr>
<tr>
<td>Knowledge Vault</td>
<td>45M</td>
<td>1.6B</td>
<td>1.1K</td>
<td>4.5K</td>
</tr>
</tbody>
</table>

* as of 2014

- It’s more than large: High heterogeneity of KBs
- *If it’s hard to write SQL on simple relational tables, it’s only harder to write SPARQL on large knowledge bases*
  - Even harder on automatically constructed KBs with a massive, loosely-defined schema
Schema-agnostic KB Querying

“Barack Obama Sr. grandchildren”

Keyword query: query like search engine

“Who are Barack Obama Sr.’s grandchildren?”

Natural language query: like asking a friend

Barack Obama Sr.

grandchildren

Graph query: add a little structure

<Barack Obama Sr., Malia Obama>

Query by example: Just show me examples
Tutorial Outline

• Introduction
• Part I: Effort–Light StructMine
  • Tea break at 3:00pm
• Part II: Schema-agnostic KB Querying
• Summary & Future Directions