Building Structured Databases of Factual Knowledge from Massive Text Corpora

Xiang Ren  Meng Jiang  Jingbo Shang  Jiawei Han

Department of Computer Science
University of Illinois at Urbana-Champaign
Turning Unstructured Text Data into Structures

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

Knowledge & Insights

(Chakraborty, 2016)
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

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

A Life Science Use Case: Identifying “Distinctively Related Entities”

Collaborate with UCLA Heart BD2K Center & Mayo Clinic

What proteins are distinctively associated with Cardiomyopathy?

Prior Art: Extracting Structures with Repeated Human Effort

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 ...
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?

**Human labeling effort**

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

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

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

- **Distantly-supervised structure discovery**
  - 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**

---

A Review of Previous Efforts
“Distant” Supervision: What Is It?

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

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

Number of Wikipedia articles

Rapidly growing!

“Un-matchable”

Text corpus

Knowledge Bases

Rapidly growing!

Human crowds

(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?

It is my favorite city in the United States

The United States needs a new strategy to meet this challenge

(Ren et al., KDD’15)
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>Accuracy of “Expansion”</td>
<td>Pick the “best” labels based on the context (for both “matchable” and “un-matchable”)</td>
</tr>
</tbody>
</table>

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

It is my favorite city in the United States

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

- Text corpus
- Data-driven text segmentation (SIGMOD’15, WWW’16)
- Entity names & context units
- Structures from the remaining unlabeled data
- Corpus-specific Structure Discovery (KDD’15, KDD’16, EMNLP’16, WWW’17)
- Partially-labeled corpus
- Knowledge bases
Effort-Light StructMine: Methodology

Text corpus

Data-driven text segmentation
(SIGMOD'15, WWW'16)

Entity names & context units

Corpus-specific Structure Discovery
(KDD’15, KDD’16, EMNLP’16, WWW’17)

Partially-labeled corpus

Knowledge bases

Pre-defined Type Structures

VS

Open Type Structures

Structures from the remaining unlabeled data

Wikipedia

Wikidata

Freebase

DBpedia

Yago

Type Structures

Type Structures

Corpus-specific Structure Discovery

KDD’15, KDD’16, EMNLP’16, WWW’17

Text segmentation
(SIGMOD’15, WWW’16)
Effort-Light StructMine: Methodology

Text corpus

Data-driven text segmentation (SIGMOD’15, WWW’16)

Entity names & context units

Corpus-specific Structure Discovery (KDD’15, KDD’16, EMNLP’16, WWW’17)

Partially-labeled corpus

Knowledge bases

Structures from the remaining unlabeled data

Typing of Entities and Relations vs Meta Pattern-Based Attribute Discovery
Tutorial Outline

• Introduction
• Part I: Quality Phrase Mining
• Part II: Joint Typing of Entities and Relations
• Part III: Meta Pattern-Based Attribute Discovery
• Summary & Future Directions