On the Power of Massive Text Data

JIAWEI HAN
COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

JANUARY 15, 2018
Outline

- From Big Data to Actionable Knowledge: How?
- On the Power of Structured Networks and Cubes
- Mining Structures from Unstructured Text
  - Phrase Mining
  - Entity/Relationship Resolution and Typing
  - MetaPAD: Meta Pattern-Directed Structure Discovery
  - Multifaceted Taxonomy Mining
- Construction of Multidimensional TextCube
- Looking Forward: Multidimensional Analysis of Massive Text Data
From Big Data to Actionable Knowledge: How?

- Ubiquity of big unstructured data
  - Big Data: Over 80% of our data is from text/natural language/social media, unstructured/semi-structured, noisy, dynamic, ..., but inter-related!

- How to turn such big data to big, actionable knowledge systematically?
  - Relying on the **magic power** of big, unstructured data
  - Three keywords:
    - **Structuring**
      - Mining hidden structures from text data
    - **Networking**
      - Turning text data into typed networks and text-cubes
    - **Mining**
      - Mining networks and cubes to generate actionable knowledge
Outline

- From Big Data to Actionable Knowledge: How?
- On the Power of Structured Networks and Cubes
- Mining Structures from Unstructured Text
  - Phrase Mining
  - Entity/Relationship Resolution and Typing
  - MetaPAD: Meta Pattern-Directed Structure Discovery
  - Multifaceted Taxonomy Mining
- Construction of Multidimensional TextCube
- Looking Forward: Multidimensional Analysis of Massive Text Data
The Power of Structured Information Networks

- Heterogeneous networks: Multiple object and link types
  - Medical network: Patients, doctors, diseases, contacts, treatments
  - Bibliographic network: Publications, authors, venues (e.g., DBLP > 2 million papers)
- PubMed is a huge heterogeneous info network
- Document links to: gene, protein, disease, drug, species, ...
- Rich knowledge can be mined from such networks

<table>
<thead>
<tr>
<th>Knowledge hidden in DBLP Network</th>
<th>Mining Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who are the <strong>leading</strong> researchers on Web search?</td>
<td>Ranking</td>
</tr>
<tr>
<td>Who are the <strong>peer</strong> researchers of Jure Leskovec?</td>
<td>Similarity Search</td>
</tr>
<tr>
<td>Whom <strong>will</strong> Christos Faloutsos <strong>collaborate with</strong>?</td>
<td>Relationship Prediction</td>
</tr>
<tr>
<td>How was the field of Data Mining <strong>emerged</strong> or <strong>evolving</strong>?</td>
<td>Network Evolution</td>
</tr>
<tr>
<td>Which authors are <strong>rather different</strong> from his/her peers in IR?</td>
<td>Outlier/anomaly detection</td>
</tr>
</tbody>
</table>
PathPredict: Meta-Path Based Relationship Prediction

- Who will be your new coauthors in the next 5 years?
- Meta path-guided prediction of links and relationships
- Philosophy: Meta path relationships among similar typed links share similar semantics and are comparable and inferable
- Co-author prediction (A—P—A) [Sun et al., ASONAM’11]
- Use topological features encoded by meta paths, e.g., citation relations between authors (A—P→P→A)

Meta-Path

<table>
<thead>
<tr>
<th>Meta-path</th>
<th>Semantic Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A \rightarrow P \rightarrow P \rightarrow A$</td>
<td>$a_i$ cites $a_j$</td>
</tr>
<tr>
<td>$A \leftarrow P \rightarrow P \rightarrow A$</td>
<td>$a_i$ is cited by $a_j$</td>
</tr>
<tr>
<td>$A \rightarrow P \rightarrow V \rightarrow P \rightarrow A$</td>
<td>$a_i$ and $a_j$ publish in the same venues</td>
</tr>
<tr>
<td>$A \rightarrow P \rightarrow A \rightarrow P \rightarrow A$</td>
<td>$a_i$ and $a_j$ are co-authors of the same authors</td>
</tr>
<tr>
<td>$A \rightarrow P \rightarrow T \rightarrow P \rightarrow A$</td>
<td>$a_i$ and $a_j$ write the same topics</td>
</tr>
<tr>
<td>$A \rightarrow P \rightarrow P \rightarrow P \rightarrow A$</td>
<td>$a_i$ cites papers that cite $a_j$</td>
</tr>
<tr>
<td>$A \leftarrow P \rightarrow P \leftarrow P \rightarrow A$</td>
<td>$a_i$ is cited by papers that are cited by $a_j$</td>
</tr>
<tr>
<td>$A \rightarrow P \rightarrow P \leftarrow P \rightarrow A$</td>
<td>$a_i$ and $a_j$ cite the same papers</td>
</tr>
<tr>
<td>$A \leftarrow P \rightarrow P \rightarrow P \rightarrow A$</td>
<td>$a_i$ and $a_j$ are cited by the same papers</td>
</tr>
</tbody>
</table>

Meta-paths between authors of length ≤ 4
Experiment on DBLP: Predicting Future Co-Authors

- Explain the prediction power of each meta-path
- Wald Test for logistic regression
- Higher prediction accuracy than using projected homogeneous network
- 11% higher in prediction accuracy

Evaluation of the prediction power of different meta-paths

<table>
<thead>
<tr>
<th>Meta Path</th>
<th>p-value</th>
<th>significance level</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A \rightarrow P \rightarrow A$</td>
<td>0.0378</td>
<td>**</td>
</tr>
<tr>
<td>$A \leftarrow P \rightarrow A$</td>
<td>0.0077</td>
<td>***</td>
</tr>
<tr>
<td>$A \rightarrow V \rightarrow P \rightarrow A$</td>
<td>1.2974e-174</td>
<td>****</td>
</tr>
<tr>
<td>$A \rightarrow P \rightarrow A$</td>
<td>1.1484e-126</td>
<td>****</td>
</tr>
<tr>
<td>$A \leftarrow P \leftarrow T \leftarrow P \rightarrow A$</td>
<td>3.4867e-51</td>
<td>****</td>
</tr>
<tr>
<td>$A \leftarrow P \rightarrow P \rightarrow P \rightarrow A$</td>
<td>0.7459</td>
<td>*</td>
</tr>
<tr>
<td>$A \rightarrow P \leftarrow P \rightarrow P \rightarrow A$</td>
<td>0.0647</td>
<td>*</td>
</tr>
<tr>
<td>$A \leftarrow P \rightarrow P \leftarrow P \rightarrow A$</td>
<td>9.7641e-11</td>
<td>****</td>
</tr>
<tr>
<td>$A \leftarrow P \leftarrow P \rightarrow P \rightarrow A$</td>
<td>0.0966</td>
<td>*</td>
</tr>
</tbody>
</table>

1 *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$; ****: $p < 0.001$


Social relations play more important role?

<table>
<thead>
<tr>
<th>Rank</th>
<th>Hybrid heterogeneous features</th>
<th># Shared authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Philip S. Yu</td>
<td>Philip S. Yu</td>
</tr>
<tr>
<td>2</td>
<td>Raymond T. Ng</td>
<td>Ming-Syan Chen</td>
</tr>
<tr>
<td>3</td>
<td>Osmar R. Zaifane</td>
<td>Divesh Srivastava</td>
</tr>
<tr>
<td>4</td>
<td>Ling Feng</td>
<td>Kotagiri Ramamohanarao</td>
</tr>
<tr>
<td>5</td>
<td>David Wai-Lok Cheung</td>
<td>Jeffrey Xu Yu</td>
</tr>
</tbody>
</table>

Co-author prediction for Jian Pei: Only 42 among 4809 candidates are true first-time co-authors! (Feature collected in [1996, 2002]; Test period in [2003, 2009])
The Power of Text Cube: Multi-Dimensional Text Analysis

- From TextCube to EventCube [KDD’13 demo]
- Keyword- or entity-based search or summary of documents
- CASeOLAP [EngBul’16]: Comparative summary/mining
Comparative Summarization in Text Cube

- A cell is located in a multi-dimensional space
- Comparative summarization makes sense
  - Given a query `<China, Economy>`
  - Target documents have frequent phrases
  - Be specific to “China”+“Economy”

- Comparing with its sibling cells in every dimension
- Criteria
  - Integrity, popularity, distinctness
  - E.g., Japanese Yen, U.S. dollars
Effectiveness of Comparative Summary on Real-World Cases

Mining distinct phrases: 2016 news data

Mining Distinct relationships between 6 subcategories of cardiovascular diseases and proteins: PubMed Abstracts
Bottleneck: Mining Unstructured Text for Structures

- One of the most challenging issue at mining big data: structuring and mining text!!
- Bottleneck: How to automatically generate structures from text data?
  - Automated mining of phrases, topics, entities, links and types from text corpora
Outline

- From Big Data to Actionable Knowledge: How?
- On the Power of Structured Networks and Cubes
- Mining Structures from Unstructured Text
  - Phrase Mining
  - Entity/Relationship Resolution and Typing
  - MetaPAD: Meta Pattern-Directed Structure Discovery
  - Multifaceted Taxonomy Mining
- Construction of Multidimensional TextCube
- Looking Forward: Multidimensional Analysis of Massive Text Data
### Phrase Mining: From Raw Corpus to Quality Phrases and Segmented Corpus

<table>
<thead>
<tr>
<th>Raw Corpus</th>
<th>Quality Phrases</th>
<th>Segmented Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ A small set of labels or a general KB</td>
<td>Integrating phrase mining with phrasal segmentation</td>
<td></td>
</tr>
</tbody>
</table>

**Input Raw Corpus**

**Phrase Mining**

**Quality Phrases**

**Segmented Corpus**

- **Document 1**
  Citation recommendation is an interesting but challenging research problem in data mining area.

- **Document 2**
  In this study, we investigate the problem in the context of heterogeneous information networks using data mining technique.

- **Document 3**
  Principal Component Analysis is a linear dimensionality reduction technique commonly used in machine learning applications.

#### Integrating phrase mining with phrasal segmentation

- **Unsupervised:** TOPMINE: A. El-Kishky, et al., Scalable Topical Phrase Mining from Text Corpora”, VLDB’15
- **Weakly supervised:** SegPhrase: J. Liu et al., Mining Quality Phrases from Massive Text Corpora. SIGMOD’15 (Grand Prize in Yelp Dataset Challenge)
- **Distantly supervised:** AutoPhrase: J. Shang, et al., Automated Phrase Mining from Massive Text Corpora, 2017
**TopMine: Frequent Pattern Mining + Statistical Analysis**

First perform frequent *contiguous pattern* mining to extract candidate phrases and their counts.

Based on significance score [Church et al.’91]:

$$\alpha(P_1, P_2) \approx (f(P_1 \bullet P_2) - \mu_0(P_1, P_2))/\sqrt{f(P_1 \bullet P_2)}$$

---

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Raw freq.</th>
<th>True freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[support vector machine]</td>
<td>90</td>
<td>80</td>
</tr>
<tr>
<td>[vector machine]</td>
<td>95</td>
<td>0</td>
</tr>
<tr>
<td>[support vector]</td>
<td>100</td>
<td>20</td>
</tr>
</tbody>
</table>

---

Markov Blanket Feature Selection for Support Vector Machines.
What Kind of Phrases Are of “High Quality”?

- Judging the quality of phrases
  - Popularity
    - “information retrieval” vs. “cross-language information retrieval”
  - Concordance
    - “powerful tea” vs. “strong tea”
    - “active learning” vs. “learning classification”
  - Informativeness
    - “this paper” (frequent but not discriminative, not informative)
  - Completeness
    - “vector machine” vs. “support vector machine”
# ToPMine: Experiments on Yelp Reviews

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>unigrams</td>
<td>coffee</td>
<td>food</td>
<td>room</td>
<td>store</td>
</tr>
<tr>
<td></td>
<td>ice</td>
<td>good</td>
<td>parking</td>
<td>shop</td>
</tr>
<tr>
<td></td>
<td>cream</td>
<td>place</td>
<td>hotel</td>
<td>prices</td>
</tr>
<tr>
<td></td>
<td>flavor</td>
<td>ordered</td>
<td>stay</td>
<td>find</td>
</tr>
<tr>
<td></td>
<td>egg</td>
<td>chicken</td>
<td>time</td>
<td>place</td>
</tr>
<tr>
<td></td>
<td>chocolate</td>
<td>roll</td>
<td>nice</td>
<td>buy</td>
</tr>
<tr>
<td></td>
<td>breakfast</td>
<td>sushi</td>
<td>place</td>
<td>selection</td>
</tr>
<tr>
<td></td>
<td>tea</td>
<td>restaurant</td>
<td>great</td>
<td>items</td>
</tr>
<tr>
<td></td>
<td>cake</td>
<td>dish</td>
<td>area</td>
<td>love</td>
</tr>
<tr>
<td></td>
<td>sweet</td>
<td>rice</td>
<td>pool</td>
<td>great</td>
</tr>
<tr>
<td>n-grams</td>
<td>ice cream</td>
<td>spring rolls</td>
<td>parking lot</td>
<td>grocery store</td>
</tr>
<tr>
<td></td>
<td>iced tea</td>
<td>food was good</td>
<td>front desk</td>
<td>great selection</td>
</tr>
<tr>
<td></td>
<td>french toast</td>
<td>fried rice</td>
<td>spring training</td>
<td>farmer’s market</td>
</tr>
<tr>
<td></td>
<td>hash browns</td>
<td>egg rolls</td>
<td>staying at the hotel</td>
<td>great prices</td>
</tr>
<tr>
<td></td>
<td>frozen yogurt</td>
<td>chinese food</td>
<td>dog park</td>
<td>parking lot</td>
</tr>
<tr>
<td></td>
<td>eggs benedict</td>
<td>pad thai</td>
<td>room was clean</td>
<td>wal mart</td>
</tr>
<tr>
<td></td>
<td>peanut butter</td>
<td>dim sum</td>
<td>pool area</td>
<td>shopping center</td>
</tr>
<tr>
<td></td>
<td>cup of coffee</td>
<td>thai food</td>
<td>great place</td>
<td>great place</td>
</tr>
<tr>
<td></td>
<td>iced coffee</td>
<td>pretty good</td>
<td>staff is friendly</td>
<td>prices are reasonable</td>
</tr>
<tr>
<td></td>
<td>scrambled eggs</td>
<td>lunch specials</td>
<td>free wifi</td>
<td>love this place</td>
</tr>
</tbody>
</table>
### SegPhrase: Weakly Supervised Quality Phrase Mining

- **SegPhrase+** (SIGMOD’15. won Grand prize of 2015 Yelp Data Set challenge, used in TripAdvisor, ...)
- Open source in Github with keyword: Jingbo Shang
- **Major features**
  - Frequent pattern mining
  - Feature extraction
  - Phrasal segmentation
  - Explore classification power, using 300 high-quality labels provided by human experts
- Interesting set of phrases generated from the titles & abstracts of SIGKDD proceedings

#### Query vs. SIGKDD

<table>
<thead>
<tr>
<th>Query</th>
<th>SIGKDD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Method</strong></td>
<td><strong>SegPhrase+</strong></td>
</tr>
<tr>
<td>1</td>
<td>data mining</td>
</tr>
<tr>
<td>2</td>
<td>data set</td>
</tr>
<tr>
<td>3</td>
<td>association rule</td>
</tr>
<tr>
<td>4</td>
<td>knowledge discovery</td>
</tr>
<tr>
<td><strong>5</strong></td>
<td><strong>time series</strong></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>51</td>
<td>association rule mining</td>
</tr>
<tr>
<td>52</td>
<td>rule</td>
</tr>
<tr>
<td><strong>53</strong></td>
<td><strong>concept drift</strong></td>
</tr>
<tr>
<td>54</td>
<td>knowledge acquisition</td>
</tr>
<tr>
<td>55</td>
<td><strong>gene expression data</strong></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>201</td>
<td>web content</td>
</tr>
<tr>
<td>202</td>
<td>frequent subgraph</td>
</tr>
<tr>
<td>203</td>
<td>intrusion detection</td>
</tr>
<tr>
<td>204</td>
<td>categorical attribute</td>
</tr>
<tr>
<td>205</td>
<td>user preference</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Only in SegPhrase+**

- **Only in Chunking**

- **Only in SegPhrase**

- **Chunking (TF-IDF & C-Value)**
AutoPhrase: Automated Phrase Mining

- Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, Jiawei Han, "AutoPhrase: Automated Phrase Mining from Massive Text Corpora", arXiv, Feb. 2017

- Automatic extraction of high-quality phrases (e.g., scientific terms and general entity names) in a given corpus (e.g., research papers and news)
  - No human efforts
  - Multiple languages
  - High performance—precision, recall, efficiency
Experiments and Performance Comparison

- **Datasets:**
  - DBLP: Scientific Paper, English, 91.6M, 618MB, 29K
  - EN: Wikipedia Article, English, 808.0M, 3.94GB, 184K
  - ES: Wikipedia Article, Spanish, 791.2M, 4.06GB, 65K
  - CN: Wikipedia Article, Chinese, 371.9M, 1.56GB, 29K

- **Comparing methods**
  - SegPhrase/WrapSegPhrae (encoding preprocessing for handling non-English)
  - TF-IDF/TextRank

<table>
<thead>
<tr>
<th>Rank</th>
<th>EN Phrase</th>
<th>CN Translation (Explanation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Elf Aquitaine</td>
<td>(the name of a soccer team)</td>
</tr>
<tr>
<td>2</td>
<td>Arnold Sommerfeld</td>
<td>Absinth</td>
</tr>
<tr>
<td>3</td>
<td>Eugene Wigner</td>
<td>(the name of a novel/TV-series)</td>
</tr>
<tr>
<td>4</td>
<td>Tarpon Springs</td>
<td>notebook computer, laptop</td>
</tr>
<tr>
<td>5</td>
<td>Sean Astin</td>
<td>Secretary of Party Committee</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20.001</td>
<td>ECAC Hockey</td>
<td>African countries</td>
</tr>
<tr>
<td>20.002</td>
<td>Sacramento Bee</td>
<td>The Left (German: Die Linke)</td>
</tr>
<tr>
<td>20.003</td>
<td>Bering Strait</td>
<td>Fraser Valley</td>
</tr>
<tr>
<td>20.004</td>
<td>Jacknife Lee</td>
<td>Hippocampus</td>
</tr>
<tr>
<td>20.005</td>
<td>WXYZ-TV</td>
<td>Mitsuki Saiga (a voice actress)</td>
</tr>
<tr>
<td>99,994</td>
<td>John Gregson</td>
<td>Computer Science and Technology</td>
</tr>
<tr>
<td>99,995</td>
<td>white-tailed eagle</td>
<td>Fonterra (a company)</td>
</tr>
<tr>
<td>99,996</td>
<td>rhombic dodecahedron</td>
<td>The Vice President of Writers</td>
</tr>
<tr>
<td>99,997</td>
<td>great spotted woodpecker</td>
<td>Association of China</td>
</tr>
<tr>
<td>99,998</td>
<td>David Manners</td>
<td>Vitamin B</td>
</tr>
</tbody>
</table>

- **Phrase Mining Results:**

![Graphs showing precision and recall for different datasets: DBLP, Yelp, EN, ES, CN]
Outline

- From Big Data to Actionable Knowledge: How?
- On the Power of Structured Networks and Cubes
- Mining Structures from Unstructured Text
  - Phrase Mining
  - Entity/Relationship Resolution and Typing
  - MetaPAD: Meta Pattern-Directed Structure Discovery
  - Multifaceted Taxonomy Mining
- Construction of Multidimensional TextCube
- Looking Forward: Multidimensional Analysis of Massive Text Data
Recognizing Typed Entities

Identifying token span as entity mentions in documents and labeling their types
— Enabling structured analysis of unstructured text corpus

The best BBQ I’ve tasted in Phoenix! I had the pulled pork sandwich with coleslaw and baked beans for lunch. ... The owner is very nice. ... The best BBQ:Food I’ve tasted in Phoenix:LOC! I had the [pulled pork sandwich]:Food with coleslaw:Food and [baked beans]:Food for lunch. ... The owner:JOB_TITLE is very nice. ...

Target Types

Plain text

Text with typed entities

Traditional methods: Expensive human labor on annotation of 500 documents for entity extraction and 20,000 queries for entity linking

Can we use the “distant labels” in the KBs?

Social media challenge!
**Entity Recognition and Typing: Challenges and Solutions**

- **Challenge 1:** Domain Restriction: Extensive training, use general-domain corpora, not work well on specific, dynamic or emerging domains (e.g., tweets, Yelp reviews)
  - Solution: Domain-agnostic phrase mining: Extracts candidate entity mentions with minimal linguistic assumption (e.g., only use POS tagging)

- **Challenge 2:** Name ambiguity: Multiple entities may share the same surface name
  - Solution: Model each mention based on its surface name and context

- **Challenge 3:** Context Sparsity: There are many ways to describe the same relation
  - Solution: cluster relation phrase, infer synonymous relation phrases

---

While Griffin is not the part of Washington’s plan on Sunday’s game, ...

... news from Washington indicates that the congress is going to...

It is one of the best state parks in Washington.

---

**Table:**

<table>
<thead>
<tr>
<th>Sentence</th>
<th>freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>The magnitude 9.0 quake caused widespread devastation in [Kesennuma city]</td>
<td>12</td>
</tr>
<tr>
<td>... tsunami that ravaged [northeastern Japan] last Friday</td>
<td>31</td>
</tr>
<tr>
<td>The resulting tsunami devastate [Japan]’s northeast</td>
<td>244</td>
</tr>
</tbody>
</table>
The ClusType Framework: Phrase Segmentation and Heterogeneous Graph Construction

- POS-constrained phrase segmentation for mining candidate entity mentions and relation phrases, simultaneously
- Construct a heterogeneous graph to represent available information in a unified form

Entity mentions are kept as individual objects **to be disambiguated**

Linked to entity surface names & relation phrases

**Weight assignment**: The more two objects are likely to share the same label, the larger the weight will be associated with their connecting edge
The Framework: Mutual Enhancement of Type Propagation and Relation Phrase Clustering

- With the constructed graph, formulate a graph-based semi-supervised learning of two tasks jointly:

  - Type propagation on heterogeneous graph
  - Multi-view relation phrase clustering

  Derived entity argument types serve as good feature for clustering relation phrases
  Propagate type information among entities bridges via synonymous relation phrases

  Mutually enhancing each other; leads to quality recognition of unlinkable entity mentions
**ClusType: Comparing with State-of-the-Art Systems**

<table>
<thead>
<tr>
<th>Methods</th>
<th>NYT</th>
<th>Yelp</th>
<th>Tweet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern (Stanford, CONLL’14)</td>
<td>0.301</td>
<td>0.199</td>
<td>0.223</td>
</tr>
<tr>
<td>SemTagger (U Utah, ACL’10)</td>
<td>0.407</td>
<td>0.296</td>
<td>0.236</td>
</tr>
<tr>
<td>NNPLB (UW, EMNLP’12)</td>
<td>0.637</td>
<td>0.511</td>
<td>0.246</td>
</tr>
<tr>
<td>APOLLO (THU, CIKM’12)</td>
<td>0.795</td>
<td>0.283</td>
<td>0.188</td>
</tr>
<tr>
<td>FIGER (UW, AAAI’12)</td>
<td>0.881</td>
<td>0.198</td>
<td>0.308</td>
</tr>
<tr>
<td>ClusType (KDD’15)</td>
<td>0.939</td>
<td>0.808</td>
<td>0.451</td>
</tr>
</tbody>
</table>

- **vs. bootstrapping**: context-aware prediction on “un-matchable”
- **vs. label propagation**: group similar relation phrases
- **vs. FIGER**: no reliance on complex feature engineering

NYT: 118k news articles (1k manually labeled for evaluation); Yelp: 230k business reviews (2.5k reviews are manually labeled for evaluation); Tweet: 302 tweets (3k tweets are manually labeled for evaluation)

Precision ($P$) = $\frac{\text{Correctly-typed mentions}}{\text{System-recognized mentions}}$, Recall ($R$) = $\frac{\text{Correctly-typed mentions}}{\text{ground-truth mentions}}$, F1 score = $\frac{2(P \times R)}{(P + R)}$
Going Deeper: Fine-Grained Entity Typing

Donald Trump is mentioned in sentences S1-S3

- Distant supervision
- Assign same types (blue region) to all the mentions
- Does not consider local contexts when assigning type labels
- Introduce label noise to the mentions

The types assigned to entity Trump include person, artist, actor, politician, businessman, while only {person, politician} are correct types for the mention “Trump” in S1
Label Noise Reduction: Framework Overview

1. Generate text features and construct a heterogeneous graph
2. Perform joint embedding of the constructed graph G into the same low-dimensional space
3. For each mention, search its candidate type sub-tree in a top-down manner and estimate the true type-path from learned embedding
CoType: Co-Embedding for Typing Entities and Relations
CoType: Comparing with State-of-the-Arts RE Systems

Given candidate relation mentions, predict its relation type if it expresses a relation of interest; otherwise, output “None”

- DeepWalk (StonyBrook, KDD’14): homogeneous graph embedding
- DS+Logistic (Stanford, ACL’09): trains logistic classifier on DS
- LINE (MSR, WWW’15): joint feature and type embedding
- MultiR (UW, ACL’11): distantly-supervised, models noisy labels
- CoType-RM (WWW’17): only models relation mentions
- CoType (WWW’17): models entity-relation interactions
Outline

- From Big Data to Actionable Knowledge: How?
- On the Power of Structured Networks and Cubes
- Mining Structures from Unstructured Text
  - Phrase Mining
  - Entity/Relationship Resolution and Typing
  - MetaPAD: Meta Pattern-Directed Structure Discovery
  - Multifaceted Taxonomy Mining
- Construction of Multidimensional TextCube
- Looking Forward: Multidimensional Analysis of Massive Text Data
MetaPAD: Meta Pattern-driven Attribute Discovery from Massive Text Corpora

- Meng Jiang, Jingbo Shang, Xiang Ren, Taylor Cassidy, Lance Kaplan, Timothy Hanratty, and Jiawei Han, “MetaPAD: Meta Pattern-driven Attribute Discovery from Massive Text Corpora”, KDD 2017

- Motivation: Given a sentence in a large corpus, “President Blaise Compaoré’s government of Burkina Faso was founded…”, ...

- We may find:
  - Blaise Compaoré: $PERSON.POLITICIAN
  - age: 65
  - Burkina Faso: $COUNTRY

- Attribute Discovery: Two tasks
  - Task 1: (entity, attribute name, attribute value)
    - (Burkina Faso, president, Blaise Compaoré)
    - (Burkina Faso, population, 17 million)
    - (Blaise Compaoré, age, 65)

  - Task 2: (entity type, attribute name)
    - ($COUNTRY, president)
    - ($COUNTRY, population)
    - ($PERSON, age)

  - Instance-level
  - Type-level
The Meta-Pattern Methodology

Generate patterns with massive instances in the data

Meta pattern segmentation

Meta patterns:
- President Blaise Compaoré’s government of Burkina Faso was founded ...
- President Barack Obama’s government of U.S. claimed that...
- U.S. President Barack Obama visited ...

Adjust types for appropriate granularity

Generate massive triples by matching the meta patterns

Joint extraction

Group synonymous patterns by massive triples

No heavy annotation required
No domain knowledge required
No query log required if we can recognize and type the entities in the same manner...

⟨Burkina Faso, {president}, Blaise Compaoré⟩
⟨U.S., {president}, Barack Obama⟩
Meta-Pattern Mining & Grouping Synonymous Patterns

Mining Meta-Patterns
• Phrase mining
• Entity typing
• Refined typing

Criteria: Quality meta-patterns
✓ Frequency
✓ Concordance: “$PERSON’s wife”
✓ Completeness: “$COUNTRY president” vs. “$COUNTRY president $POLITICIAN”
✓ Informativeness: “$PERSON and $PERSON” vs. “$PERSON’s wife, $PERSON”

Finding synonymous patterns

⟨$COUNTRY, president, $POLITICIAN⟩

president

$COUNTRY president $POLITICIAN

⟨United States, Barack Obama⟩

⟨$PERSON, {age, -year-old}, $DIGIT⟩

⟨Barack Obama, 55⟩

⟨$PERSON, $DIGIT, $PERSON’s age is $DIGIT⟩

⟨Justin Trudeau, 43⟩

⟨$PERSON, a $DIGIT-year-old⟩
## Patterns, Entities and Attribute Values Found in News Corpus

<table>
<thead>
<tr>
<th>Meta patterns</th>
<th>Entity</th>
<th>Attribute value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$COUNTRY President $POLITICIAN</td>
<td>United States</td>
<td>Barack Obama</td>
</tr>
<tr>
<td>$COUNTRY’s president $POLITICIAN</td>
<td>Russia</td>
<td>Vladimir Putin</td>
</tr>
<tr>
<td>President $POLITICIAN of $COUNTRY</td>
<td>France</td>
<td>Francois Hollande</td>
</tr>
<tr>
<td>…</td>
<td>COUNTRY's</td>
<td>politician's government of COUNTRY</td>
</tr>
<tr>
<td>$POLITICIAN’s government of $COUNTRY</td>
<td>Burkina Faso</td>
<td>Blaise Compaoré</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Meta patterns</th>
<th>Entity</th>
<th>Attribute value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$COMPANY CEO $PERSON</td>
<td>Apple</td>
<td>Tim Cook</td>
</tr>
<tr>
<td>$COMPANY chief executive $PERSON</td>
<td>Facebook</td>
<td>Mark Zuckerberg</td>
</tr>
<tr>
<td>$PERSON, the $COMPANY CEO,</td>
<td>Hewlett-Packard</td>
<td>Carly Fiorina</td>
</tr>
<tr>
<td>…</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$COMPANY former CEO $PERSON</td>
<td>Infor</td>
<td>Charles Phillips</td>
</tr>
<tr>
<td>$PERSON, the $COMPANY former CEO,</td>
<td>Afghan Citadel</td>
<td>Roya Mahboob</td>
</tr>
</tbody>
</table>
Patterns/Entities/Values Found in Medical Science Corpus

<table>
<thead>
<tr>
<th>Meta patterns</th>
<th>Entity</th>
<th>Attribute value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Treatment was used to treat } \text{Disease} $\text{Disease using the } \text{Treatment}</td>
<td>zoledronic acid therapy</td>
<td>Paget’s disease of bone</td>
</tr>
<tr>
<td>$\text{Treatment has been used to treat } \text{Disease} $\text{Treatment of patients with } \text{Disease}</td>
<td>bisphosphonates</td>
<td>osteoporosis</td>
</tr>
<tr>
<td></td>
<td>calcitonin</td>
<td>Paget’s disease of bone</td>
</tr>
<tr>
<td></td>
<td>calcitonin</td>
<td>osteoporosis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{Bacteria was resistant to } \text{Antibiotics} $\text{Bacteria are resistant to } \text{Antibiotics}</td>
<td>corynebacterium striatum BM4687</td>
<td>gentamicin</td>
</tr>
<tr>
<td>$\text{Bacteria is the most resistant to } \text{Antibiotics} $\text{Bacteria, particularly those resistant to } \text{Antibiotics}</td>
<td>corynebacterium striatum BM4687</td>
<td>tobramycin</td>
</tr>
<tr>
<td></td>
<td>methicillin-susceptible S aureus</td>
<td>vancomycin</td>
</tr>
<tr>
<td></td>
<td>multidrug-resistant enterobacteriaceae</td>
<td>gentamicin</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outline

- From Big Data to Actionable Knowledge: How?
- On the Power of Structured Networks and Cubes
- Mining Structures from Unstructured Text
  - Phrase Mining
  - Entity/Relationship Resolution and Typing
  - MetaPAD: Meta Pattern-Directed Structure Discovery
  - Multifaceted Taxonomy Mining
- Construction of Multidimensional TextCube
- Looking Forward: Multidimensional Analysis of Massive Text Data
Automated construction of topic taxonomy

- Selected method: **spherical clustering**—Use **embeddings** to find semantically consistent clusters
  - Domain-specific terms can be clustered together
    - “machine learning”, “learning algorithm”, ...
  - Where do the general terms go?
    - “computer science”, “method”, “paper”, ...

Diagram:
- Documents
- Topic Dimension
- “computer science”
- “graphics”
- “machine learning”

Recursive construction
- computer_science
  - data_minning
  - networking
  - computer_vision
  - machine_learning
  - information_retrieval
  - clustering
  - reinforcement_learning
  - deep_learning
  - bayesian_network
  - classification
- CS
  - CG
  - ML
  - IR
- Clustering
- Classification
Adaptive Spherical Clustering

- Design a ranking module to select **representative phrases** for each cluster
  - Conduct comparative analysis (combining **popularity** and **concentration**)
  - Does this phrase better fit my cluster or my siblings’?
- Push the **background phrases** back to the general node
  - “computer science”, “paper” → the higher-level node (root node)
  - “machine learning”, “ml”, “classification” → the “ML” node
- The set of remaining phrases leads to more separable clustering
Local Embedding vs. Global Embedding

- Global embedding (embedding learning on the global dataset) does not work
- Terms at different granularity can have close embeddings
- Ex. “Information Extraction”: similar to “text mining”, “NLP”, “machine learning”
- Solution: local embedding:
  - For each “sub-topic” node, learn local embedding only on relevant documents
  - Only preserve information relevant to the “sub-topic”
**TaxonGen: Adaptive Spherical Clustering + Local Embedding**

- Phrase mining + Adaptive spherical clustering: Generate top-level clusters
- Local embedding: Generate lower level clusters

Experiment with the DBLP dataset

High quality multi-level hierarchy generated automatically
Outline

- From Big Data to Actionable Knowledge: How?
- On the Power of Structured Networks and Cubes
- Mining Structures from Unstructured Text
  - Phrase Mining
  - Entity/Relationship Resolution and Typing
  - MetaPAD: Meta Pattern-Directed Structure Discovery
  - Multifaceted Taxonomy Mining
- Construction of Multidimensional TextCube
- Looking Forward: Multidimensional Analysis of Massive Text Data
Cube Construction: Which Document Goes to Which Cell?

- Cell-based Document Allocation
- Which document goes to which cell?

Dimensions
- USA
- China
- Japan
- Economy
- Sports
- Politics
- 2014
- 2015
- 2016

Documents
- Sports
- Politics
- Economy

Corpus

<table>
<thead>
<tr>
<th>ID</th>
<th>Document Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>... The super bowl is on air from Chicago, Illinois. The NFL has decided that best coach of 2017 is from...</td>
</tr>
<tr>
<td>2</td>
<td>... make a speech in Shanghai that economy plan is to make sure manufactory industry of China...</td>
</tr>
<tr>
<td>3</td>
<td>... in Dec 2015, attacks continued in France for two more days, taking the lives of six others</td>
</tr>
</tbody>
</table>

Text Cube

Dimensions
- Topic Dimension
- Time Dimension
- Location Dimension
How to Put Documents into the Right Cube Cell?

- Major challenges on putting docs into the right cell
  - Few would like label the “training sets”
  - So many cells, so many documents
  - Dimension values are often “under-represented”
    - E.g., Topic dimension: Sports, economy, politics, ...
  - Documents are often “over-represented” on single dimension
  - Ex. “… The super bowl is on air from Chicago, Illinois. The NFL has decided that best coach of 2017 is from …
- Our methodology: Dimension-aware joint embedding
  - Constructing an L-T-D (label-term-document) graph
Dimension-Aware Joint Embedding

Joint Embedding

Document Focalization

Label Expansion

Embed everything in the same semantic space
Constructing Text Cubes with Massive Data, Few Labels

- Dimension focusing—**Dimension-Focal Score**, a discriminative measure
- A term \( t \) is “focal” to dimension \( L \)
- The documents with \( t \) has very imbalanced labels (KL-divergence can be a good measure)
- Ex.

- Label expansion: Combining two measures for seed expansion
  - Discriminativeness
  - Using focal score
  - Popularity
  - Example:

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Label</th>
<th>1st Expansion</th>
<th>2nd Expansion</th>
<th>3rd Expansion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Movies</td>
<td>films</td>
<td>director</td>
<td>hollywood</td>
<td></td>
</tr>
<tr>
<td>Baseball</td>
<td>inning</td>
<td>hits</td>
<td>pitch</td>
<td></td>
</tr>
<tr>
<td>Tennis</td>
<td>wimbledon</td>
<td>french open</td>
<td>grand slam</td>
<td></td>
</tr>
<tr>
<td>Business</td>
<td>company</td>
<td>chief executive</td>
<td>industry</td>
<td></td>
</tr>
<tr>
<td>Law Enforcement</td>
<td>litigation</td>
<td>law</td>
<td>county courthouse</td>
<td></td>
</tr>
<tr>
<td><strong>Location</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brazil</td>
<td>brazilian</td>
<td>sao paulo</td>
<td>confederations cup</td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td>sydney</td>
<td>australian</td>
<td>melbourne</td>
<td></td>
</tr>
<tr>
<td>Spain</td>
<td>madrid</td>
<td>barcelona</td>
<td>la liga</td>
<td></td>
</tr>
<tr>
<td>China</td>
<td>chinese</td>
<td>shanghai</td>
<td>beijing</td>
<td></td>
</tr>
</tbody>
</table>
Experiment: Overall Document Allocation Performance

- **Data:**
  - 3,080 NYT articles
  - 2 dimensions: topic, location

- **Baselines**
  - **IR:** tf-idf score with label query
  - **IR + QE:** expand query with word2vec similarity
  - **Topic Model (TM)**
  - **Word2vec & Word2vec + DF**
  - **Dataless Classification**

- **Ablations**
  - **Joint-Emb:** use representations from joint embedding
  - **JE+DF:** joint embedding with document focalization
  - **JE+LE:** joint embedding with label expansion

---

**Table:**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th><strong>Topic Dimension</strong></th>
<th><strong>Location Dimension</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Micro-F1</td>
<td>Macro-F1</td>
</tr>
<tr>
<td><strong>IR</strong></td>
<td>0.3963</td>
<td>0.4520</td>
</tr>
<tr>
<td><strong>IR+QE</strong></td>
<td>0.4112</td>
<td>0.4744</td>
</tr>
<tr>
<td><strong>Word2vec</strong></td>
<td>0.5928</td>
<td>0.3890</td>
</tr>
<tr>
<td><strong>Word2vec+DF</strong></td>
<td>0.6101</td>
<td>0.3980</td>
</tr>
<tr>
<td><strong>Topic Model</strong></td>
<td>0.6264</td>
<td>0.3620</td>
</tr>
<tr>
<td><strong>Dataless</strong></td>
<td>0.5882</td>
<td>0.3724</td>
</tr>
<tr>
<td><strong>Joint-Emb</strong></td>
<td>0.6938</td>
<td>0.4992</td>
</tr>
<tr>
<td><strong>JE+DF</strong></td>
<td>0.7862</td>
<td>0.5235</td>
</tr>
<tr>
<td><strong>JE+LE</strong></td>
<td>0.7347</td>
<td>0.5081</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>0.7957</td>
<td>0.5413</td>
</tr>
</tbody>
</table>
Outline

- From Big Data to Actionable Knowledge: How?
- On the Power of Structured Networks and Cubes
- Mining Structures from Unstructured Text
  - Phrase Mining
  - Entity/Relationship Resolution and Typing
  - MetaPAD: Meta Pattern-Directed Structure Discovery
  - Multifaceted Taxonomy Mining
- Construction of Multidimensional TextCube
- Looking Forward: Multidimensional Analysis of Massive Text Data
Looking Forward: MD Analysis of Massive Text Data

- From big data to big knowledge
- A key problem: Multidimensional mining massive text data
- Lots to be explored!!!
Our Journey: From Big Data to Big Structures & Knowledge

Han, Kamber and Pei, Data Mining, 3rd ed. 2011
Yu, Han and Faloutsos (eds.), Link Mining, 2010
Sun and Han, Mining Heterogeneous Information Networks, 2012
Y. Sun: SIGKDD’13 Dissertation Award
Wang and Han, Mining Latent Entity Structures, 2015
C. Wang: SIGKDD’15 Dissertation Award
Acknowledgements

Thanks for the research support from: ARL/NSCTA, NIH, NSF, DARPA, DTRA, ......
References

- A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han, “Scalable Topical Phrase Mining from Text Corpora”, VLDB’15
- M. Jiang, J. Shang, X. Ren, T. Cassidy, L. Kaplan, T. Hanratty and J. Han, “MetaPAD: Meta Pattern-driven Attribute Discovery from Massive Text Corpora”, KDD’17
- J. Liu, J. Shang, C. Wang, X. Ren, J. Han, “Mining Quality Phrases from Massive Text Corpora”, SIGMOD’15
- J. Liu, J. Shang, J. Han, Phrase Mining from Massive Text and Its Applications, Morgan & Claypool, 2017
- X. Ren, A. El-Kishky, C. Wang, F. Tao, C. R. Voss, H. Ji and J. Han, “ClusType: Effective Entity Recognition and Typing by Relation Phrase-Based Clustering”, KDD’15
- X. Ren, Z. Wu, W. He, M. Qu, C. Voss, H. Ji, T. Abdelzaher, J. Han, "CoType: Joint Extraction of Typed Entities and Relations with Knowledge Bases", WWW'17
- X. Ren, et al., "LifeNet: A Structured Network-Based Knowledge Exploration and Analytics System for Life Sciences", ACL'17 (system demo), July 2017