Entity Matching across Heterogeneous Sources

Yang Yang*, Yizhou Sun†, Jie Tang*, Bo Ma#, and Juanzi Li*

*Tsinghua University  †Northeastern University  #Carnegie Mellon University
Apple Inc. VS Samsung Co.

- A patent infringement suit starts from 2012.
  - Lasts 2 years, involves $158+ million and 10 countries.
  - 7 out of 35546 patents are involved.

How to find patents relevant to a specific product?

<table>
<thead>
<tr>
<th></th>
<th>Galaxy S II Skyrocket</th>
<th>Galaxy S III</th>
<th>Galaxy Tab 2 (10.1)</th>
<th>Stratosphere</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple’s patent</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

SAMSUNG devices accused by APPLE.
Cross-Source Entity Matching

- Given an entity in a **source** domain, we aim to find its matched entities from **target** domain.
  - Product-patent matching;
  - Cross-lingual matching;
  - Drug-disease matching.
Problem

Source 1: Siri's Wiki page

iOS iPhone iPod iPad
intelligent personal assistant
Cydia knowledge navigator
voice control Apple server
natural language user interface

Source 2: Patents

heuristic algorithms
distribution system
speech recognition
data source text-to-speech

Method for improving voice recognition
Universal interface for retrieval of information in a computer system
search engine descriptors
object relevant area ranking module rank candidate
synchronize database host device media
customized processor graphical user interface

Input 1: Dual source corpus

\{C_1, C_2\}, where C_t=\{d_1, d_2, \ldots, d_n\}
is a collection of entities

Input 2: Matching relation matrix

L_{ij} =
\begin{cases}
1, & \text{d}_i \text{ and } \text{d}_j \text{ are matched} \\
0, & \text{not matched} \\
?, & \text{unknown}
\end{cases}
Challenges

Two domains have *less or no overlapping* in content

Daily expression

VS

Professional expression
Challenges

1. Two domains have *less or no overlapping* in content

2. How to model the topic-level relevance probability
Our Approach

Cross-Source Topic Model
Baseline

1. Topic extraction

Wikipedia

C1

$C_1$

$d_1$

$d_2$

$\ldots$

$d_n$

USPTO

C2

$C_2$

$z_1$

$z_3$

$z_4$

$\ldots$

$d'_m$

query

Little-overlapping content

$\rightarrow$ disjoint topic space

2. Ranking candidates by topic similarity

\[
Sim(d_1, d_2) = \frac{p_{d_1} \cdot p_{d_2}}{|p_{d_1}| \times |p_{d_2}|}
\]
Cross-Sampling

How latent topics influence matching relations?

If C=0, sample topics according to the topic distribution of $d_m$.

Bridge topic space by leveraging known matching relations.

d_n is matched with $d_m$.
Inferring Matching Relation

Infer matching relations by leveraging extracted topics.

\[ \rho(l_d, d'_t = 1|z_d, z_d', \gamma) \propto \exp[\gamma^T(z_d \odot z_{d'})] \]
Cross-Source Topic Model

Step 1:

Step 2:

Latent topics ← Matching relations
Model Learning

• Variational EM
  – Model parameters: \( \{\varphi, \gamma\} \)
  – Variational parameters: \( \{\vartheta, \tau, \eta, \epsilon\} \)
  – E-step:
    
    \[
    \begin{align*}
    \eta_{d,c} &= \beta_{d,c} + N_d \times \epsilon_{d,c} \\
    \tau_{d,k} &= \alpha_k + \sum_{n=1}^{N_d} \vartheta_{d,n,k} \\
    \epsilon_{d,n,c} &\propto \exp\{\Psi(\eta_{d,c}) - \Psi(\sum_{i \in R(d)} \eta_{d,i})\} \\
    \vartheta_{d,n,k} &\propto \sum_{d' \in (R(d), d)} (\exp\{\sum_{d'' \neq d'} \frac{\gamma_k \sum_{i=1}^{N_{d''}} \theta_{d'',i,k}}{N_{d''} N_{d'}}\} + \Psi(\tau_{d',k}) - \Psi(\sum_{j=1}^{K} \tau_{d',j})\}) \epsilon_{d,n,d'} \times \varphi_{t,k,v}
    \end{align*}
    \]

  – M-step:
    
    \[
    \begin{align*}
    \varphi_{t,k,v} &\propto \sum_{d=1}^{D_t} \sum_{n=1}^{N_d} \vartheta_{d,n,k} 1(w^t_{d,n} = v) \\
    \gamma_k &= \frac{\sum_{d,d'} 1}{2 \sum_{d,d'} l_{d,d'}[(\bar{Y}_d - \bar{Y}_{d'}) \circ (\bar{Y}_d - \bar{Y}_{d'})]_k}
    \end{align*}
    \]
Experiments

Task I: Product-patent matching
Task II: Cross-lingual matching
Task I: Product-Patent Matching

• Given a Wiki article describing a product, finding all patents relevant to the product.

• Data set:
  – 13,085 Wiki articles;
  – 15,000 patents from USPTO;
  – 1,060 matching relations in total.
**Experimental Results**

**Training**: 30% of the matching relations randomly chosen.

<table>
<thead>
<tr>
<th>Method</th>
<th>P@3</th>
<th>P@20</th>
<th>MAP</th>
<th>R@3</th>
<th>R#20</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS+LDA</td>
<td>0.111</td>
<td>0.083</td>
<td>0.109</td>
<td>0.011</td>
<td>0.046</td>
<td>0.053</td>
</tr>
<tr>
<td>RW+LDA</td>
<td>0.111</td>
<td>0.117</td>
<td>0.123</td>
<td>0.033</td>
<td>0.233</td>
<td>0.429</td>
</tr>
<tr>
<td>RTM</td>
<td>0.501</td>
<td>0.233</td>
<td>0.416</td>
<td>0.057</td>
<td>0.141</td>
<td>0.171</td>
</tr>
<tr>
<td>RW+CST</td>
<td>0.667</td>
<td>0.167</td>
<td>0.341</td>
<td>0.200</td>
<td>0.333</td>
<td>0.668</td>
</tr>
<tr>
<td>CST</td>
<td>0.667</td>
<td>0.250</td>
<td>0.445</td>
<td>0.171</td>
<td>0.457</td>
<td>0.683</td>
</tr>
</tbody>
</table>

**Content Similarity based on LDA (CS+LDA)**: cosine similarity between two entities’ topic distribution extracted by LDA.

**Random Walk based on LDA (RW+LDA)**: random walk on a graph where edges indicate the hyperlinks between Wiki articles and citations between patents.

**Relational Topic Model (RTM)**: used to model links between documents.

**Random Walk based on CST (RW+CST)**: uses CST instead of LDA comparing with RW+LDA.
Task II: Cross-lingual Matching

• Given an English Wiki article, we aim to find a Chinese article reporting the same content.

• Data set:
  – 2,000 English articles from Wikipedia;
  – 2,000 Chinese articles from Baidu Baike;
  – Each English article corresponds to one Chinese article.
# Experimental Results

**Training:** 3-fold cross validation

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Measure</th>
<th>F2-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title Only</td>
<td>1.000</td>
<td>0.410</td>
<td>0.581</td>
<td>0.465</td>
</tr>
<tr>
<td>SVM-S</td>
<td>0.957</td>
<td>0.563</td>
<td>0.709</td>
<td>0.613</td>
</tr>
<tr>
<td>LFG</td>
<td>0.661</td>
<td>0.820</td>
<td>0.732</td>
<td>0.782</td>
</tr>
<tr>
<td>LFG+LDA</td>
<td>0.652</td>
<td>0.805</td>
<td>0.721</td>
<td>0.769</td>
</tr>
<tr>
<td>LFG+CST</td>
<td>0.682</td>
<td><strong>0.849</strong></td>
<td><strong>0.757</strong></td>
<td><strong>0.809</strong></td>
</tr>
</tbody>
</table>

**Title Only:** only considers the (translated) title of articles.

**SVM-S:** famous cross-lingual Wikipedia matching toolkit.

**LFG:** mainly considers the structural information of Wiki articles.

**LFG+LDA:** adds content feature (topic distributions) to LFG by employing LDA.

**LFG+CST:** adds content feature to LFG by employing CST.

### Topics Relevant to Apple and Samsung

(Topic titles are hand-labeled)

<table>
<thead>
<tr>
<th>Title</th>
<th>Top Patent Terms</th>
<th>Top Wiki Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gravity Sensing</td>
<td>Rotational, gravity, interface, sharing, frame, layer</td>
<td>Gravity, iPhone, layer, video, version, menu</td>
</tr>
<tr>
<td>Touchscreen</td>
<td>Recognition, point, digital, touch, sensitivity, image</td>
<td>Screen, touch, iPad, os, unlock, press</td>
</tr>
<tr>
<td>Application Icons</td>
<td>Interface, range, drives, icon, industrial, pixel</td>
<td>Icon, player, software, touch, screen, application</td>
</tr>
</tbody>
</table>
Prototype System
competitor analysis @ http://pminer.org

1. Electrical computers
2. Static information
3. Information storage
4. Data processing
5. Active solid-state devices
6. Computer graphics processing
7. Molecular biology and microbiology
8. Semiconductor device manufacturing

Radar Chart: topic comparison

Basic information comparison: #patents, business area, industry, founded year, etc.
Conclusion

• Study the problem of entity matching across heterogeneous sources.

• Propose the cross-source topic model, which integrates the topic extraction and entity matching into a unified framework.

• Conduct two experimental tasks to demonstrate the effectiveness of CST.
Thank You!

Entity Matching across Heterogeneous Sources

Yang Yang*, Yizhou Sun†, Jie Tang*, Bo Ma‡, and Juanzi Li*

*Tsinghua University  †Northeastern University  ‡Carnegie Mellon University

Apple Inc. VS Samsung Co.

• A patent infringement lawsuit starts from 2012.
  – Lasts over 2 years, involves $158+ million.

• How to find patents relevant to a specific product?
Problem

• Given an entity in a source domain, we aim to find its matched entities from target domain.
  – Given a textural description of a product, finding related patents in a patent database.
  – Given an English Wiki page, finding related Chinese Wiki pages.
  – Given a specific disease, finding all related drugs.
Basic Assumption

• For entities from different sources, their matching relations and hidden topics are influenced by each other.

• How to leverage the known matching relations to help link hidden topic spaces of two sources?
Cross-Sampling

d₁ and d₂ are matched …

Source 1

\[ \beta_{d_1} = (4,1,0) \]

Source 2

\[ \beta_{d_2} = (1,4,0) \]

\[ \beta_{d_3} = (0,0,4) \]

Topics

\[ Z_1 \]
\[ Z_2 \]
\[ Z_3 \]
\[ Z_4 \]
\[ Z_5 \]
Sample a new term $w_1$ for $d_1$

Toss a coin $c$, if $c=0$, sample $w_1$’s topic according to $d_1$
Cross-Sampling

3

Sample a new term $w_1$ for $d_1$
Otherwise sample $w_1$’s topic according to $d_2$

Source 1

$\beta_{d_1} = (4, 1, 0)$

Source 2

$\beta_{d_2} = (1, 4, 0)$

$\beta_{d_3} = (0, 0, 4)$

Word $w_1$

Topics

$z_1$

$z_2$

$z_3$

$z_4$

$z_5$
Parameter Analysis

(a) Number of topics $K$

(b) Ratio

(c) Precision

(d) Convergence analysis
Experimental Results

**Training**: 30% of the matching relations randomly chosen.

<table>
<thead>
<tr>
<th>Method</th>
<th>P@3</th>
<th>P@20</th>
<th>MAP</th>
<th>R@3</th>
<th>R#20</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS+LDA</td>
<td>0.111</td>
<td>0.083</td>
<td>0.109</td>
<td>0.011</td>
<td>0.046</td>
<td>0.053</td>
</tr>
<tr>
<td>RW+LDA</td>
<td>0.111</td>
<td>0.117</td>
<td>0.123</td>
<td>0.033</td>
<td>0.233</td>
<td>0.429</td>
</tr>
<tr>
<td>RTM</td>
<td>0.501</td>
<td>0.233</td>
<td>0.416</td>
<td>0.057</td>
<td>0.141</td>
<td>0.171</td>
</tr>
<tr>
<td>RW+CST</td>
<td><strong>0.667</strong></td>
<td>0.167</td>
<td>0.341</td>
<td><strong>0.200</strong></td>
<td>0.333</td>
<td>0.668</td>
</tr>
<tr>
<td>CST</td>
<td><strong>0.667</strong></td>
<td><strong>0.250</strong></td>
<td><strong>0.445</strong></td>
<td>0.171</td>
<td><strong>0.457</strong></td>
<td><strong>0.683</strong></td>
</tr>
</tbody>
</table>

**Content Similarity based on LDA (CS+LDA)**: cosine similarity between two articles’ topic distribution extracted by LDA.

**Random Walk based on LDA (RW+LDA)**: random walk on a graph where edges indicate the hyperlinks between Wiki articles and citations between patents.

**Relational Topic Model (RTM)**: used to model links between documents.

**Random Walk based on CST (RW+CST)**: uses CST instead of LDA comparing with RW+LDA.
**Experimental Results**

**Training**: 30% of the matching relations randomly chosen.

<table>
<thead>
<tr>
<th>Method</th>
<th>P@3</th>
<th>P@20</th>
<th>MAP</th>
<th>R@3</th>
<th>R#20</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS+LDA</td>
<td>0.111</td>
<td>0.083</td>
<td>0.109</td>
<td>0.011</td>
<td>0.046</td>
<td>0.053</td>
</tr>
<tr>
<td>RW+LDA</td>
<td>0.111</td>
<td>0.117</td>
<td>0.123</td>
<td>0.033</td>
<td>0.233</td>
<td>0.429</td>
</tr>
<tr>
<td>RTM</td>
<td>0.501</td>
<td>0.233</td>
<td>0.416</td>
<td>0.057</td>
<td>0.141</td>
<td>0.171</td>
</tr>
<tr>
<td>RW+CST</td>
<td><strong>0.667</strong></td>
<td>0.167</td>
<td>0.341</td>
<td><strong>0.200</strong></td>
<td>0.333</td>
<td>0.668</td>
</tr>
<tr>
<td>CST</td>
<td><strong>0.667</strong></td>
<td><strong>0.250</strong></td>
<td><strong>0.445</strong></td>
<td>0.171</td>
<td><strong>0.457</strong></td>
<td><strong>0.683</strong></td>
</tr>
</tbody>
</table>

**Content Similarity based on LDA (CS+LDA)**: cosine similarity between two articles’ topic distribution extracted by LDA.

**Random Walk based on LDA (RW+LDA)**: random walk on a graph where edges indicate the hyperlinks between Wiki articles and citations between patents.

**Relational Topic Model (RTM)**: used to model links between documents.

**Random Walk based on CST (RW+CST)**: uses CST instead of LDA comparing with RW +LDA.