Verification of POI and Location Pairs via Weakly Labeled Web Data

Hsiu-Min Chuang, Chia-Hui Chang
National Central University
Taoyuan, Taiwan
Agenda

- Introduction
- Problem definition
- Method
- Experiment
- Related work
- Conclusion and future work
Introduction (Background)

- **Google Maps** have replaced the past paper maps and telephone books.
  - Accommodate an unlimited number of POIs (point-of-interest)
  - The popularity of mobile devices and wireless network
Introduction (Trends)

- A market research of 5000 persons by *comScore* in 2014, 90% of users have used a **local search**.
  
  **Top1.** finding an address/location  
  **Top2.** finding a business with products/services needed  
  **Top3.** querying the phone number of a business
Terminologies

- According to the **scope** and **granularity**, 
  1. **Location**: a centroid (a pair of a longitude and latitude in a widely adopted system). **It does not change over time**.
  2. **Place**: human construct; **coarse** level of spatial granularity. Larger scale administrative constructs (i.e., city, neighborhoods. A place may contain multiple POIs.
- **POI**: human construct; a **fine** level of spatial granularity. Some attributes (i.e., name, current location, address, category). **A POI has a loose coupling with a location**.
- Focus on the POI relations between the **location (address)** and the **name (store, organization, or building)**
Motivations

• To provide local searches on maps, researches focus on deriving spatial context or geographic entities from the Web.

• **Maintenance of the crawled POI becomes a challenge.**
  – Verify whether outdated or existing POIs with time passing

• Some POI-relations may change due to grand-opening, moving, renaming and closing of business.

• An address maps to multiple stores or a store maps to multiple addresses
  – They could be either mostly correct or mostly wrong
POI coupling: Address-to-Name Mapping

- **1-to-m**: Mostly outdated
  - The same address maps to multiple store names

- **m-to-1**: Mostly correct
  - The same store name maps to multiple addresses

There are five outdated POIs in Zhupiter website

There are five *SUBWAY* branch stores in Taoyuan from Google Maps
Verification of POI Relations

- Problem definition
  - Given a POI pair (address-to-POI name), determine if the pair is outdated.

- The label is T if the pair is correct (existing), and F if the pair is outdated.

  \[
  \text{Pair}(\text{address, name}) \quad \text{Features} \quad \text{label}\{T, F\}
  \]

- The POI-relation verification problem can be regarded as a classification problem.

- Weakly-labeled data from the Web as features
How to Solve the Verification Problem?

- **Existing Methods:** Users feedback and manual verification
  - **Costly** to maintain these enormous POIs
  - **Slow progress** for updating pairs by crowd-sourcing

- **Our Method:** Verification via weakly labeled Web data
  - Labels of the training instances could be *implicit* or *noisy*
  - Web data provides *evidence* for the relation of a POI pair when the amount of the related pages is enough.
    - Collect the relevant webpages by search engines
Weakly-Labeled Web Data
- The search results of search engines

- Use Google search engine to collect features for classification
- Use address, store name, and the combination as queries, respectively.
- For each query,
  1. # of relevant results
  2. # of co-occurs in the same snippets
  3. Most recently published date
  4. Snippets similarity
  5. Ranks
The Reason of Using These Features

• Indicators of a strong POI relation:
  - The larger number of search results for POI relation
  - The larger conditional probability for finding the store name given the address query or vice versa
  - Most recently published date
  - Cosine similarity between the search snippet vectors of the address and store name queries
  - The NDCG score of store name rank from the top ten search result of the address query
## Feature Expression

Given query search results from Google using address \( a \), store \( s \) and \( a+s \), there are five kinds of features as follows.

<table>
<thead>
<tr>
<th>id</th>
<th>Name</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \log C(a) )</td>
<td># of search results for query ( a ) in log scale</td>
</tr>
<tr>
<td>2</td>
<td>( \log C(s) )</td>
<td># of search results for query ( s ) in log scale</td>
</tr>
<tr>
<td>3</td>
<td>( \log C(s,a) )</td>
<td># of search results for query ( a+s ) in log scale</td>
</tr>
<tr>
<td>4</td>
<td>( R(a+s</td>
<td>a) )</td>
</tr>
<tr>
<td>5</td>
<td>( R(a+s</td>
<td>s) )</td>
</tr>
<tr>
<td>6</td>
<td>( P(a+s</td>
<td>T_a) )</td>
</tr>
<tr>
<td>7</td>
<td>( P(a+s</td>
<td>T_s) )</td>
</tr>
<tr>
<td>8</td>
<td>( P(a+s</td>
<td>T_{a+s}) )</td>
</tr>
<tr>
<td>9</td>
<td>( \text{NDCG}(s</td>
<td>T_a) )</td>
</tr>
<tr>
<td>10</td>
<td>( \text{NDCG}(a</td>
<td>T_s) )</td>
</tr>
<tr>
<td>11</td>
<td>( \cos(T_a,T_s) )</td>
<td>the cosine similarity for snippet ( T_a ) and ( T_s )</td>
</tr>
<tr>
<td>12</td>
<td>( D(a+s) )</td>
<td>Today-( D_{a+s} ) in log scale</td>
</tr>
</tbody>
</table>
Feature Distribution

• **Yellow Page**
  – Features of outdated POIs are similar to existing POIs
  – A single feature is difficult to distinguish the POIs whether existing or outdated.
Experimental Dataset and Measures

- **1-to-m Yellow Page**
  Crawled from hiPage and iPeen
  - Manually label (6,640)
  - Unlabeled data (50,000)

- The label is denoted by the actual class
  - F: outdated pairs; T: existing pairs

- For **outdated** pairs,
  \[ A = \# \text{ of pairs that are predicted as F} \]
  \[ B = \# \text{ of pairs that are labeled as F} \]

- **Precision** \( F \) = \( \frac{A \cap B}{A} \)
- **ACC** = \( \frac{\# \text{ of pairs that are predicted correctly}}{\# \text{ of pairs}} \)
- **Recall** \( F \) = \( \frac{A \cap B}{B} \)
- **F-measure** \( F \) = \( 2 \frac{P \times R}{P + R} \)
Performance for 1-to-m Yellow Page

1. Supervised Methods

- 75% labeled data for training and 25% for testing
- Use four methods to conduct three-fold cross validation
- In terms of $F_1$, libSVM performs best
- In terms of $ACC$, Bagging performs best

<table>
<thead>
<tr>
<th>Methods</th>
<th>ACC</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBF Network</td>
<td>0.551</td>
<td>0.551</td>
<td>0.653</td>
<td>0.598</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>0.577</td>
<td>0.572</td>
<td>0.709</td>
<td>0.632</td>
</tr>
<tr>
<td>libSVM</td>
<td>0.590</td>
<td>0.585</td>
<td>0.695</td>
<td>0.635</td>
</tr>
<tr>
<td>Bagging</td>
<td>0.607</td>
<td>0.607</td>
<td>0.655</td>
<td>0.630</td>
</tr>
</tbody>
</table>
2. Semi-Supervised Method

- Concept of Tri-training

Labeled Data $L$ (May contain noise)

Train $h_i$, $h_j$, $h_k$ by $L$

Classifier $h_i$

Classifier $h_j$

Classifier $h_k$

Unlabeled Data $U$

Co-Labeling Results at $t$-th round $L_i^t$

Re-Train By $L \cup L_i^t$

Draw?

no

yes
How to obtain three classifiers

- **Approach 1**: Resample 3 different datasets
- **Approach 2**: Use 3 different features sets
- **Approach 3**: Use 3 different learning algorithms

- Training:
  - L: 60% labeled data
  - U: 50,000 unlabeled data
- Testing: 40% labeled data
- Repeat three times
Tri-training for Different Combinations

- Approaches 1 and 3 are improved, but the performance of approach 2 is reduced.
- Tri-Training iterates to reduce the error rate
  - The increasing of accuracy is not significant
  - The major improvement is for outdated examples

<table>
<thead>
<tr>
<th>Tri-Training</th>
<th>75% amount</th>
<th>75% features</th>
<th>SVM-DT-BAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Accuracy</td>
<td>.606</td>
<td>.569</td>
<td>.618</td>
</tr>
<tr>
<td>Final Accuracy</td>
<td>.607</td>
<td>.568</td>
<td>.620</td>
</tr>
<tr>
<td>Initial F1</td>
<td>.561</td>
<td>.604</td>
<td>.653</td>
</tr>
<tr>
<td>Final F1</td>
<td>.649</td>
<td>.540</td>
<td>.695</td>
</tr>
</tbody>
</table>

Tri-training with Data Diversity (SVM)

<table>
<thead>
<tr>
<th>#L+Lnew</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>8314</td>
<td>12186</td>
<td>19297</td>
<td>31512</td>
<td>48328</td>
<td></td>
</tr>
<tr>
<td>ACC</td>
<td>0.598</td>
<td>0.608</td>
<td>0.603</td>
<td>0.602</td>
<td>0.601</td>
</tr>
<tr>
<td>Error</td>
<td>0.392</td>
<td>0.331</td>
<td>0.037</td>
<td>0.014</td>
<td>0.006</td>
</tr>
</tbody>
</table>
Reason Analysis

(1) Different distribution of the training & testing dataset
- The decreased error rate in the training set doesn’t always imply the improved accuracy of the testing set.

(2) Co-labeling without confidence control
- When both classifiers have low confidence, incorrect examples may be added for training of the 3rd classifier
- This problem does not occur when classifiers have high accuracy
Distribution of **Confidence** for Different Classifiers

- **SVM**: Most examples have probability less than 0.65
- **C4.5** and **Bagging**: the probability distribution is more balanced.
- If SVM is used for tri-training, we set confidence threshold around 0.55~0.65 for tri-training.
Tri-training for 1-to-m Yellow-Page
- Performance of Different Confidence Threshold

- We set confidence 0.55~0.65 for SVM-DT-Bagging.
- When the confidence=0.65, the F1 measure can reach to 0.702.
- Selecting instance with high confidence is important.

<table>
<thead>
<tr>
<th>Tri-Training Threshold</th>
<th>0.55</th>
<th>0.60</th>
<th>0.65</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Accuracy</td>
<td>.618</td>
<td>.618</td>
<td>.618</td>
</tr>
<tr>
<td>Final Accuracy</td>
<td>.620</td>
<td>.616</td>
<td>.610</td>
</tr>
<tr>
<td>Initial F1</td>
<td>.653</td>
<td>.653</td>
<td>.653</td>
</tr>
<tr>
<td>Final F1</td>
<td>.660</td>
<td>.701</td>
<td>.702</td>
</tr>
</tbody>
</table>

Learning Curve for Unlabeled Data by Tri-Training with Confidence Selection
Related Work

- **Construction of POI DB**
  - Where the streets have no name: experiences in GIR for a developing country, *GIR*, 2013.

- **Crowdsourcing Data Refinement**

- **Semi-Supervised Learning**
Conclusions and Future Works

• We apply supervised and semi-supervised learning methods for detecting outdated POI by weakly labeled Web data.
  – For 1-to-m Yellow Pages pairs, tri-training can improve F1-measure from 0.66 to 0.702.

Future work

• Combine social network dataset and semantic features to improve the performance
• The verification task for general pages