Verification of POI and Location Pairs via Weakly Labeled Web Data

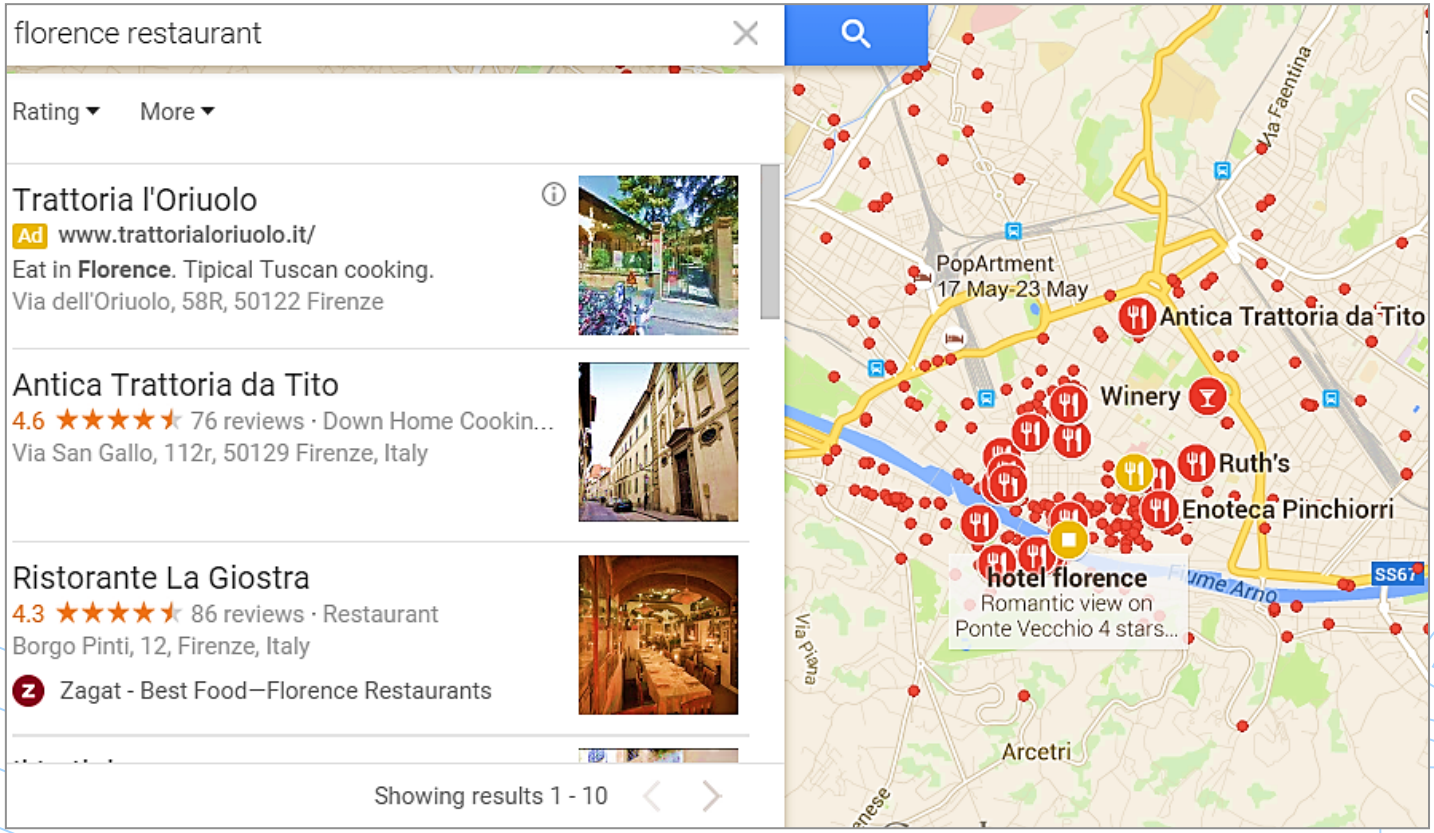
Hsiu-Min Chuang, Chia-Hui Chang
National Central University
Taoyuan, Taiwan

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



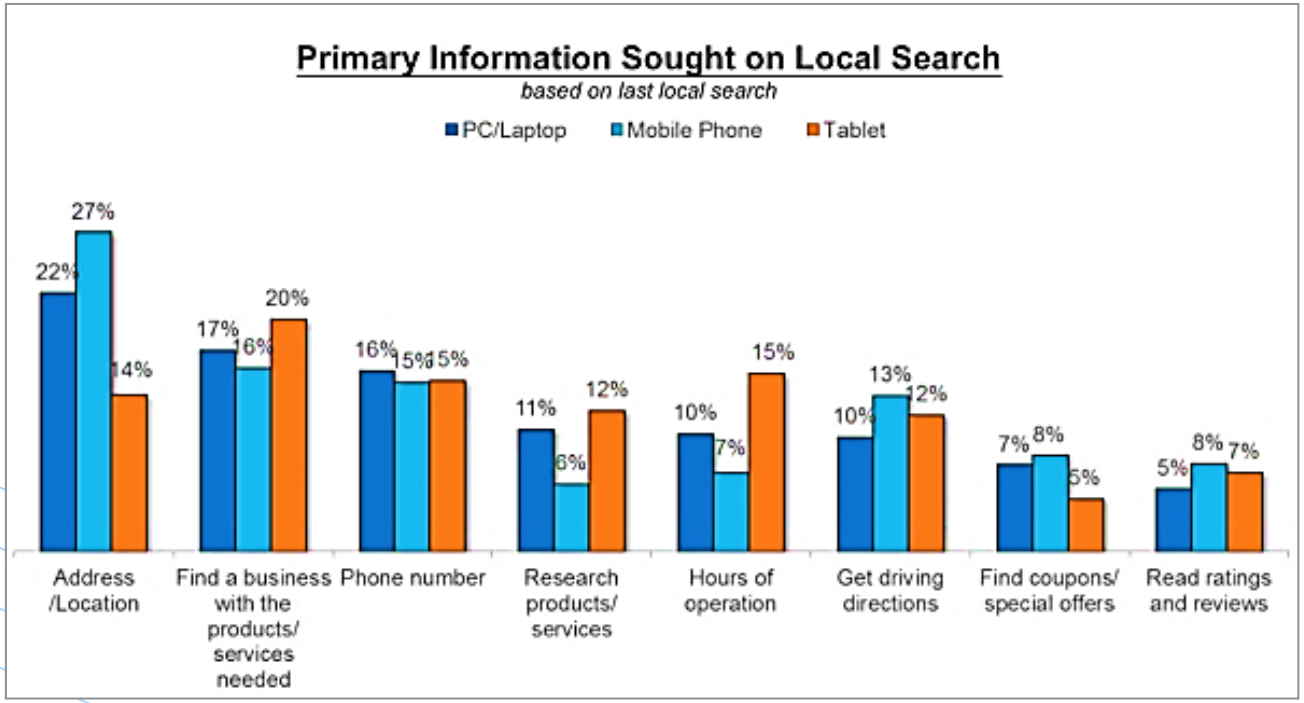
Introduction (Trends)

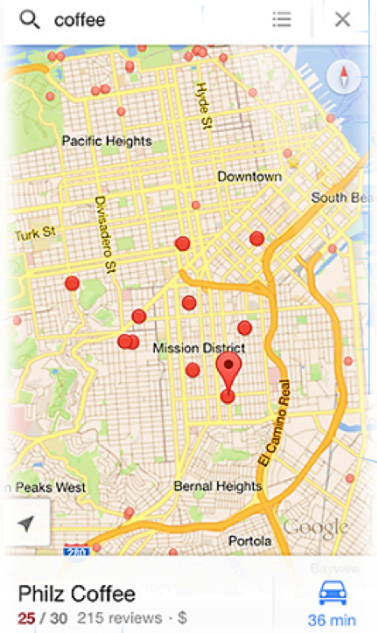
- A market research of 5000 persons by *comScore* in 2014, 90% of users have used a **local search**.

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  - Verify whether outdated or existing POIs with time passing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.

Pair(address, name)	Features	label{T, F}

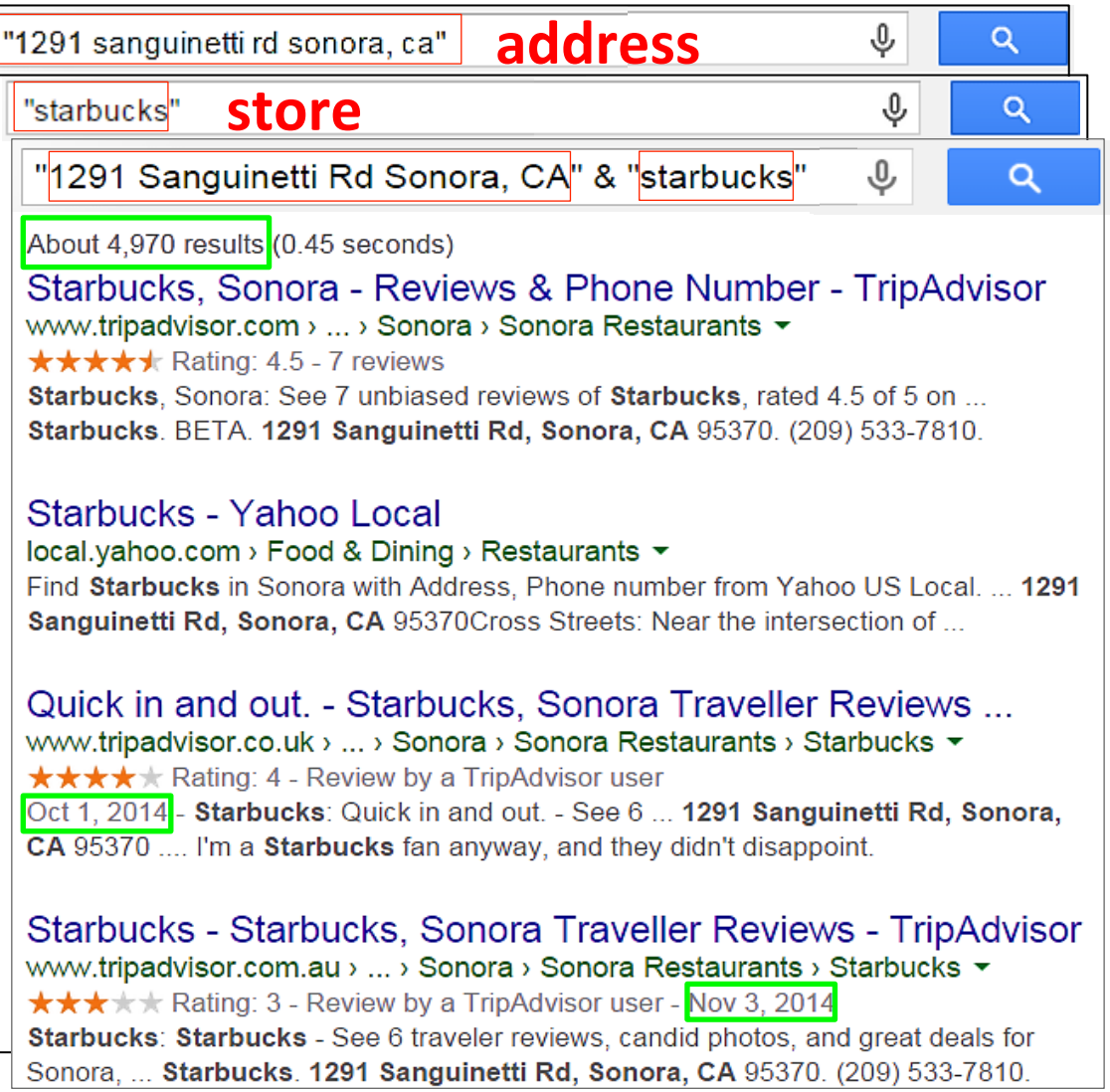
- The POI-relation verification problem can be regarded as a classification problem.
- Weakly-labeled data from the Web as features

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

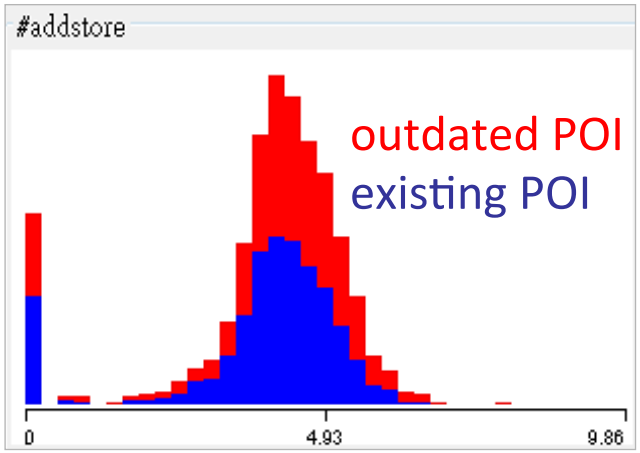
Given query search results from Google using address a, store s and a+s, there are five kinds of features as follows.

id	Name	Descriptions
1	log *C(a)*	# of search results for query a in log scale
2	log *C(s)*	# of search results for query s in log scale
3	log *C(s,a)*	# of search results for query a+s in log scale
4	R(*a+s*/*a*)	the ratio of C(a+s) to C(a)
5	R(*a+s*/*s*)	the ratio of C(a+s) to C(s)
6	P(*a+s*/T_a)	the percentage of top 10 snippets from q_a that support (a,s)
7	P(*a+s*/T_s)	the percentage of top 10 snippets from q_s that support (a,s)
8	P(*a+s*/T_{a+s})	the percentage of top 10 snippets from q_{a+s} that support (a,s)
9	NDCG(*s*/T_a)	the rank of *s* in top 10 snippets from T_a
10	NDCG(*a*/T_s)	the rank of *a* in top 10 snippets from T_s
11	cos(T_a,T_s)	the cosine similarity for snippet T_a and T_s
12	D(a+s)	Today-D_{a+s} in log scale

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.

# of search results for *a* and *s*

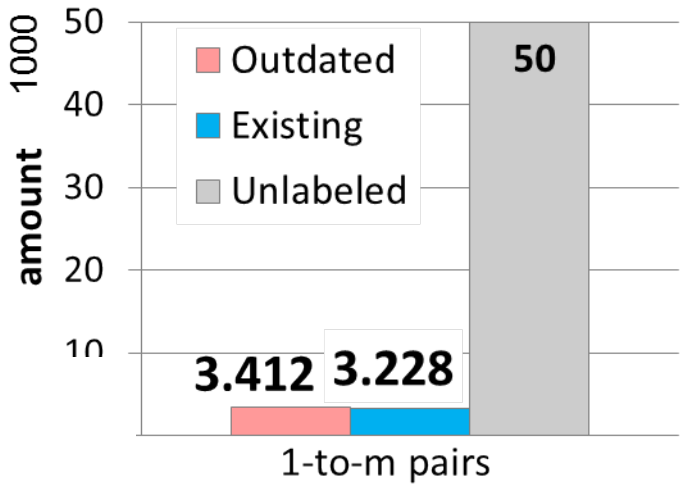


log(#add & store)

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 = # of pairs that are predicted as F

  B = # of pairs that are labeled as F

- $\text{Precision}_F = \frac{A \cap B}{A}$
- $\text{Recall}_F = \frac{A \cap B}{B}$
- $\text{F-measure}_F = 2 * \frac{P * R}{P + R}$
- $\text{ACC} = \frac{\text{\# of pairs that are predicted correctly}}{\text{\# of pairs}}$

Performance for 1-to-m Yellow Page

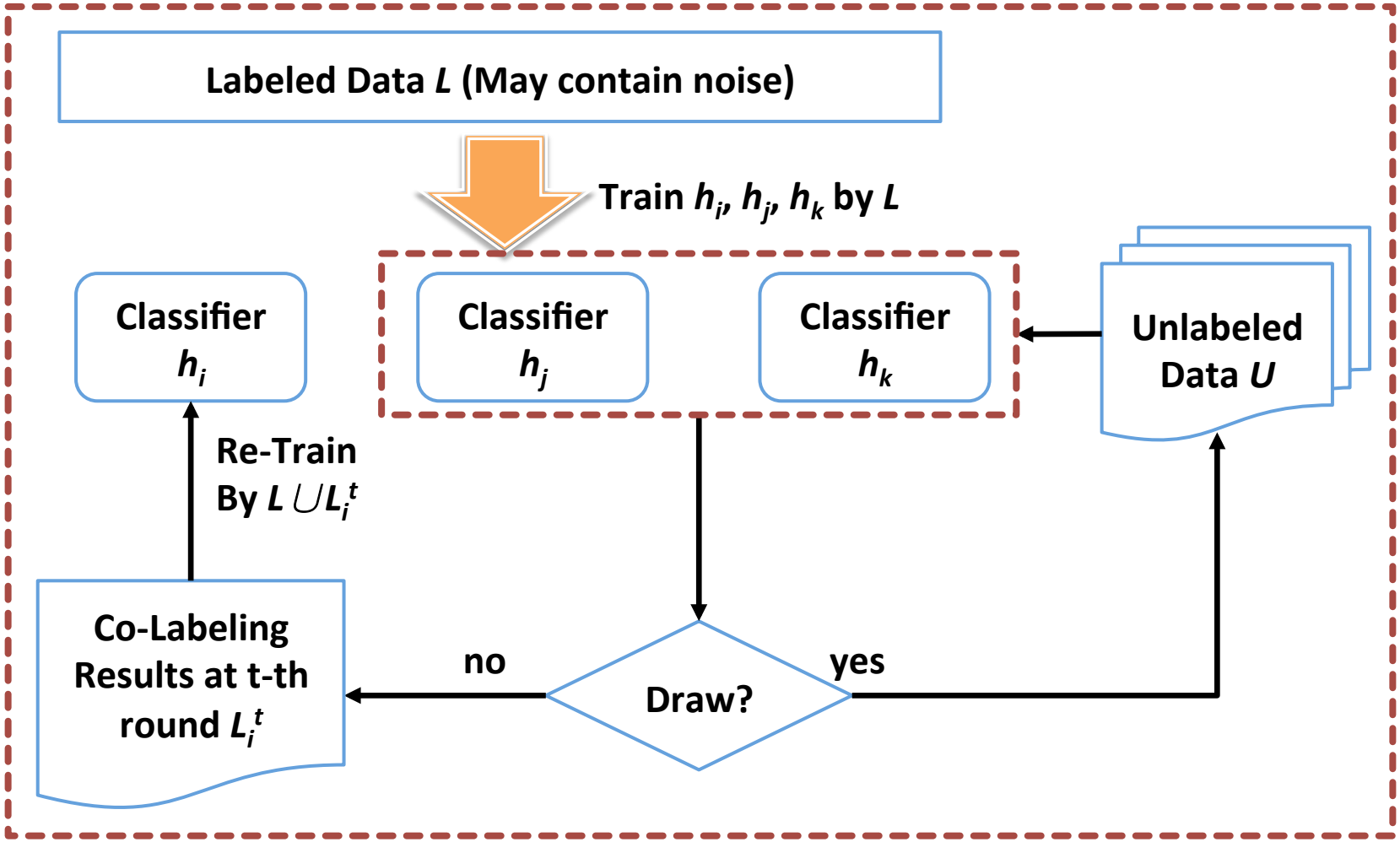
1. Supervised Methods

- 75% labeled data for training and 25% for testing
- Use four methods to conduct three-fold cross validation
- In terms of F1, libSVM performs best
- In terms of ACC, Bagging performs best

Methods	ACC	P	R	F1
RBF Network	0.551	0.551	0.653	0.598
AdaBoost	0.577	0.572	0.709	0.632
libSVM	**0.590**	0.585	0.695	**0.635**
Bagging	**0.607**	0.607	0.655	**0.630**

2. Semi-Supervised Method

- Concept of Tri-training



How to obtain three classifiers

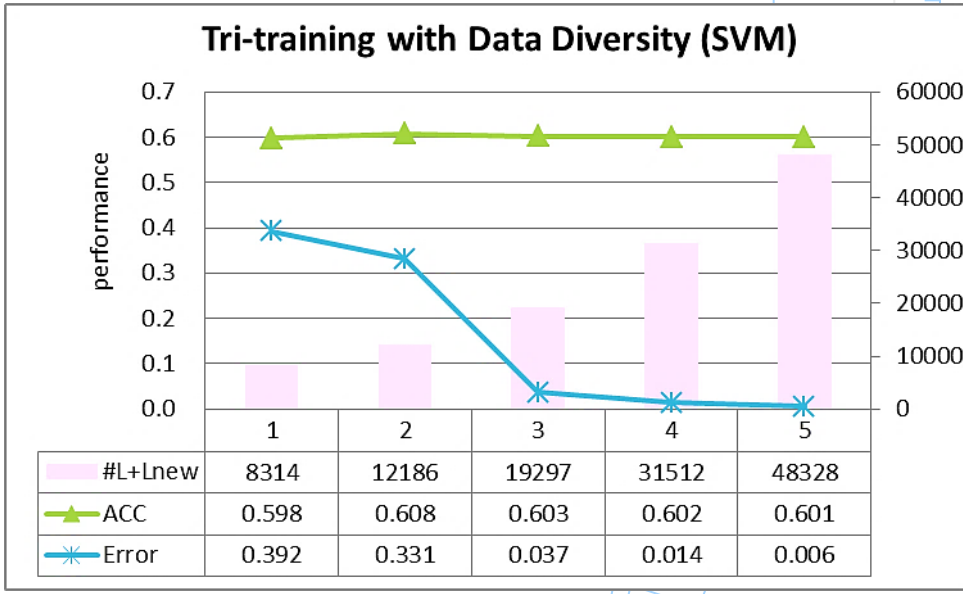
- Approach 1: Resample 3 different datasets
- Approach 2: Use 3 different features sets
- Approach 3: Use 3 different learning algorithms

- 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

Tri-Training	75% amount	75% features	SVM-DT-BAG
Initial Accuracy	.606	.569	.618
Final Accuracy	.607	.568	.620
Initial F1	.561	.604	.653
Final F1	**.649**	**.540**	**.695**



Reason Analysis

(1) Different distribution of the training & testing dataset

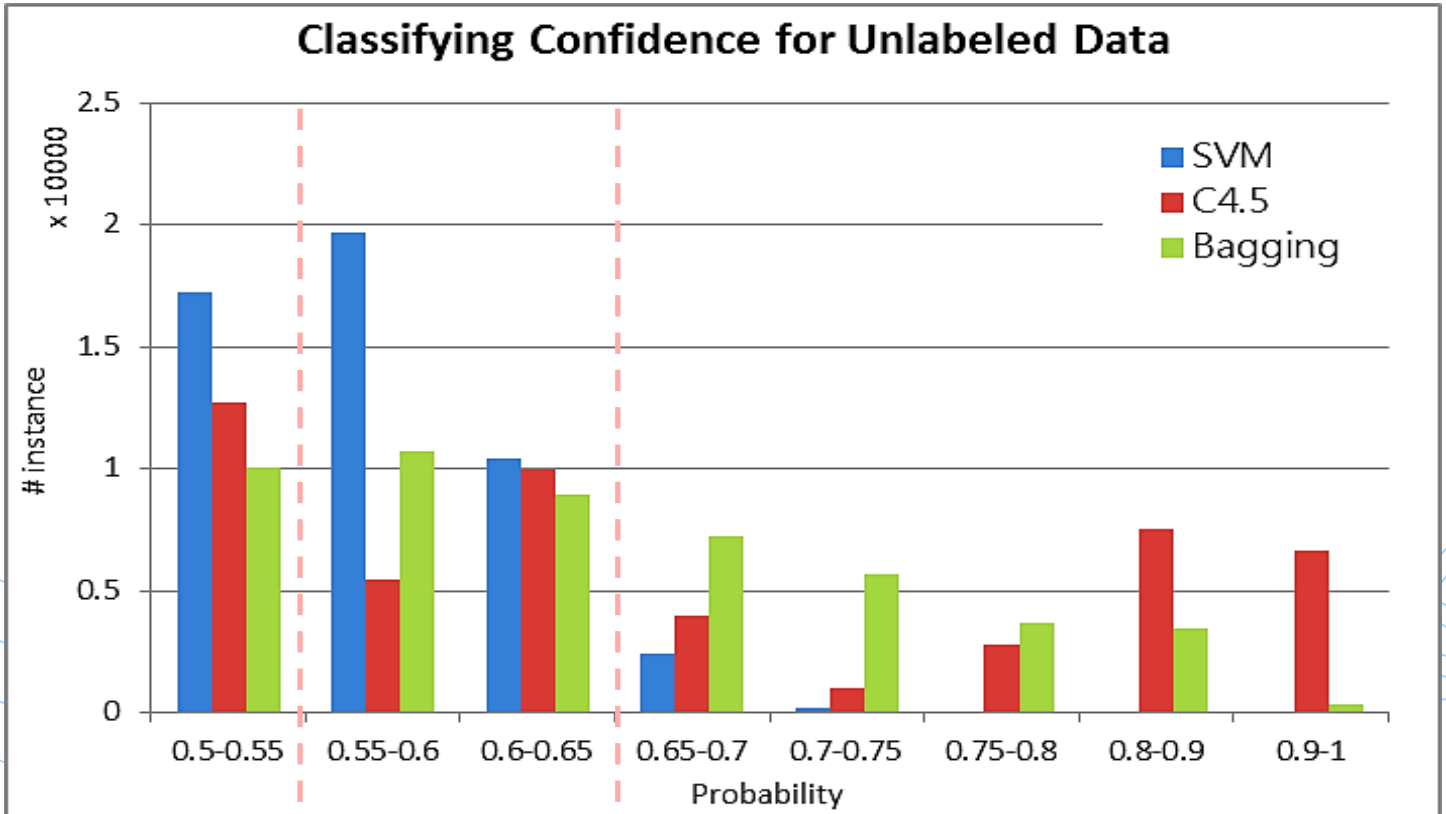
- The decreased error rate in the training set doesn't always imply the improved accuracy of the testing set.

(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 balance
- If SVM is used for tri-training, we set confidence threshold around 0.55~0.65 for tri-training.

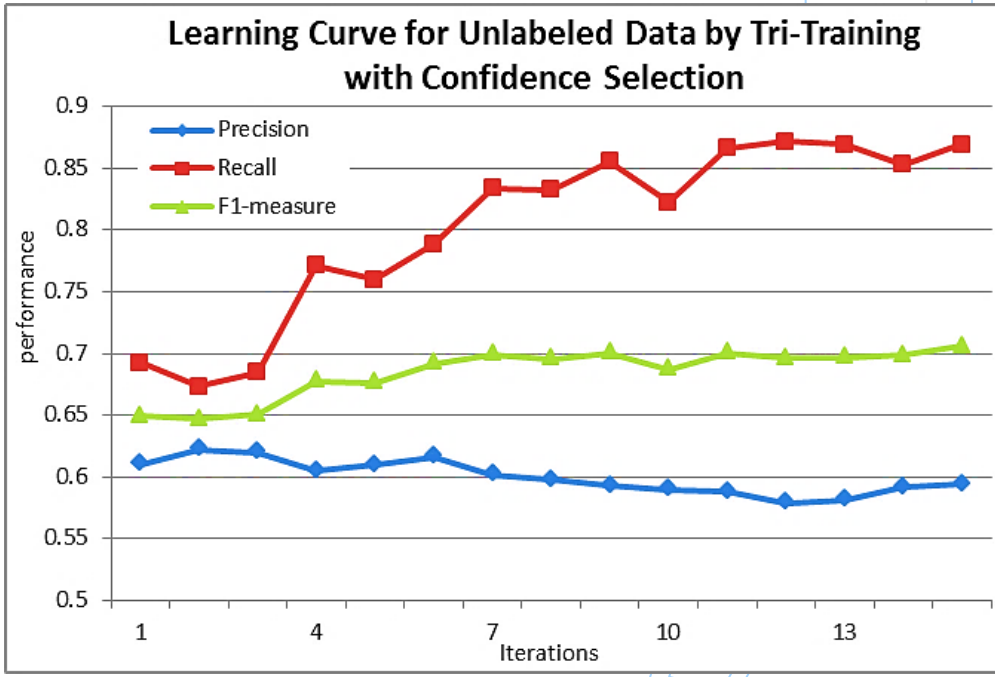


Tri-training for 1-to-m Yellow-Page

- 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.

Tri-Training Threshold	0.55	0.60	0.65
Initial Accuracy	.618	.618	.618
Final Accuracy	**.620**	.616	.610
Initial F1	.653	.653	.653
Final F1	.660	.701	**.702**



Related Work

- Construction of POI DB
  - ✓ Effective Web Crawling for Chinese Addresses and Associated Information, *EC-Web*, 2014.
  - ✓ Where the streets have no name: experiences in GIR for a developing country, *GIR*, 2013.
- Crowdsourcing Data Refinement
  - ✓ Data Quality Assurance for Volunteered Geographic Information, *WI*, 2014.
  - ✓ Deduplicating a Places Database, *WWW*, 2014.
- Semi-Supervised Learning
  - ✓ Tri-Training: Exploiting Unlabeled Data Using Three Classifiers, *TKDE*, 2005.

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