Learning an Accurate Entity Resolution Model from Crowdsourced Labels

Jingjing Wang  
Graduate School of Information Science and Technology  
Hokkaido University  
Sapporo 060-0814, Japan  
jingjing.wang@live.cn

Satoshi Oyama  
Graduate School of Information Science and Technology  
Hokkaido University  
Sapporo 060-0814, Japan  
oyama@ist.hokudai.ac.jp

Masahito Kurihara  
Graduate School of Information Science and Technology  
Hokkaido University  
Sapporo 060-0814, Japan  
kurihara@ist.hokudai.ac.jp

Hisashi Kashima  
Department of Mathematical Informatics  
The University of Tokyo  
Tokyo, Japan / JST PRESTO  
kashima@mist.i.u-tokyo.ac.jp

ABSTRACT
We investigated the use of supervised learning methods that use labels from crowd workers to resolve entities. Although obtaining labeled data by crowdsourcing can reduce time and cost, it also brings challenges (e.g., coping with the variable quality of crowd-generated data). First, we evaluated the quality of crowd-generated labels for actual entity resolution data sets. Then, we evaluated the prediction accuracy of two machine learning methods that use labels from crowd workers: a conventional LPP method using consensus labels obtained by majority voting and our proposed method that combines multiple Laplacians directly by using crowdsourced data. We discussed the relationship between the accuracy of workers’ labels and the prediction accuracy of the two methods.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—clustering; I.5.3 [Pattern Recognition]: Clustering—algorithms, similarity measures

General Terms
Algorithms, Experimentation

Keywords
Entity resolution, Crowdsourcing, Link prediction, Dimensionality reduction

1. INTRODUCTION

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

IMCOM (ICUIMC)’14 January 9-11, 2014, Siem Reap, Cambodia.  
Copyright 2014 ACM 978-1-4503-2644-5 ...$15.00.
Since she began climbing in 1988, Abby has excelled at just about every kind of climbing imaginable. In 1996, she and fellow Australian Vera Wong set the...
can be defined as an adjacency matrix:

\[ A_{ij} = \begin{cases} 1 & \text{if } x_i \text{ and } x_j \text{ have a link,} \\ 0 & \text{otherwise.} \end{cases} \]

We use consensus labels obtained by MV as the adjacency matrix. The problem can be rewritten as a generalized eigenvector problem:

\[ W^* = \arg \min_W \text{tr} \left( W^t \Phi^t L \Phi W \right) \]

s. t. \( W^t \Phi^t D^t \Phi W^t = I_d, \)

where \( \Phi \) is the design matrix defined as \( \Phi = [x_1, \ldots, x_N]^t. \) \( D \) is a diagonal degree matrix in which the entries are column sums of \( A. D_{ii} = \sum_j A_{ij}. \) \( L \) is the Laplacian matrix defined by \( L = D - A. \) \( I_d \) is the \( d \times d \) identity matrix. We use the constraint to avoid the trivial solution \((W = 0)\) and ensure the uniqueness of the solution.

Solving this optimization problem is equivalent to solve the following generalized eigenvalue problem.

\[ \Phi L \Phi^t w = \lambda \Phi D \Phi^t w. \]

By finding \( d \) eigenvectors with the smallest positive eigenvalues we can get the optimal linear projection matrix \( W^*. \)

### 2.2 Combination of Multiple Laplacians

Though we can solve entities with aggregated crowdsourced labels, the information on the distribution of workers’ judgments is missing in the learning process, e.g., the consensus label obtained by MV is truncated to 0 or 1. We introduce a combination of multiple Laplacians to entity resolution to directly derive an accurate model from crowd-generated labels. The process of finding the optimal projection matrix can be seen as

\[ W^* = \arg \min_W \sum_{t=1}^{T} \sum_{i,j} A_{ij}^{(t)} ||Wx_i - Wx_j||^2, \]

where the adjacency matrix \( A^{(t)} \) is defined as \( A^{(t)} = \{A_{ij}^{(t)}\}, \) which is provided by worker \( t. \) \( T \) is the number of crowd workers.

The problem can be also formulated as an optimization problem:

\[ W^* = \arg \min_W \sum_{t=1}^{T} \sum_{i,j} r_t A_{ij}^{(t)} ||Wx_i - Wx_j||^2, \]

where \( L^t \) is the Laplacian matrix corresponding to adjacency matrix \( A^{(t)}, \) which is defined the same as \( A \) in the LPP method. \( r_t = 1 \) if \( x_i \) and \( x_j \) have a link otherwise \( r_t = 0. \)

#### 2.2.1 Weighting Workers

The model derived for a more qualified worker has a higher degree of accuracy. To compensate for the variability in the accuracy of the crowd workers and thereby get a more accurate true model, we impose an additional weighting measure on the learning process: we assign higher weights to more accurate workers.

As a result, the process of finding the optimal projection matrix can be rewritten as

\[ W^* = \arg \min_W \sum_{t=1}^{T} \sum_{i,j} r_t A_{ij}^{(t)} ||Wx_i - Wx_j||^2, \]

where \( r_t \) is the weighting parameter.

The problem can also be formulated as an optimization problem:

\[ W^* = \arg \min_W \text{tr} \left( W^t \Phi^t \sum_{t=1}^{T} r_t L^t \Phi W \right). \]

This means that the problem can be reduced to a generalized eigenvalue problem:

\[ \Phi \sum_{t=1}^{T} r_t L^t \Phi^t w = \lambda \Phi \Phi^t w. \]

The optimal projection matrix \( W^* \) is obtained by finding the \( d \) eigenvectors with the smallest positive eigenvalues. After the data objects are mapped to the \( d \)--dimensional space by using projection matrix \( W^*, \) the links are identified on the basis of the distance between them, and the prediction accuracy can be calculated by comparing the predicted result with the gold standard.

To find the \( d \) smallest eigenvalues can be problematic, so we found the eigenvectors for the largest positive eigenvalues of the generalized eigenvalue problem by setting \( \lambda = 1 \setminus \mu. \) Then, the final generalized eigenvalue problem becomes:

\[ \Phi \sum_{t=1}^{T} r_t L^t \Phi^t w = \mu \Phi \Phi^t w. \]

Three weighting measures were used to represent crowd worker credibility. One values the weight of each worker as the degree of similarity, one uses mutual dependency principles to score the weight of each worker and without weighting. Without weighting, the labels provided by the crowd workers would be treated equally; that is, \( r_t = 1/T. \)

**Similarity.**

If we assume that most workers give true labels, a worker is more qualified if the labels provided by the worker are more similar to the labels provided by other workers, so a higher weight should be given to that worker. The similarity between labels \( A_{ij} \) provided by worker \( t, \) and labels \( A_{ij} \) provided by worker \( t, j \) is defined as

\[ \text{Sim}_{ti, tj} = \frac{\langle A_{ti}, A_{tj} \rangle}{||A_{ti}|| ||A_{tj}||}. \]

Thus, the weight of worker \( t, j \) is defined as the sum of the similarities:

\[ r_t = \sum_{j \neq i} \text{Sim}_{ti, tj}, \]

where \( T \) is the set of crowd workers, \( T = \{1, \ldots, T\}. \) In practice, we use normalization to make the sum of the weights of all workers equal to 1.

**Mutual Dependence.**

A worker-labeled page pair can be modeled as a bipartite graph in which the label given by a worker to a pair of Web pages can be seen as a directed edge that carries the “opinion” of a source node towards a target node. From the perspective of the target node (a pair of Web pages), the edge can be identified as “agreeing” or “disagreeing” by determining whether the opinion carried by the edge agrees with the majority voting opinion for the target node. In our case, if most workers label a pair of Web pages \( p \) as “0,” worker \( t, j \) labels it as “1,” and worker \( t, k \) labels it as “0,” edge \( (t, j), (t, k) \) is a disagreeing edge, and edge \( (t, j), (t, k) \) is an agreeing edge.

We use the mutual dependency principles proposed by Dai et al. [5] to evaluate of crowd worker credibility. These principles include the positive and negative mutual dependencies. An agreeing
Table 1: Training data and test data (Web pages) for each person name.

<table>
<thead>
<tr>
<th>Name</th>
<th>Training data</th>
<th>Test data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abby Watkins</td>
<td>60</td>
<td>52</td>
</tr>
<tr>
<td>Alvin Cooper</td>
<td>40</td>
<td>47</td>
</tr>
<tr>
<td>Armando Valencia</td>
<td>30</td>
<td>29</td>
</tr>
<tr>
<td>David Lodge</td>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td>Jerry Hobbs</td>
<td>40</td>
<td>48</td>
</tr>
<tr>
<td>Michael Howard</td>
<td>40</td>
<td>48</td>
</tr>
<tr>
<td>Patrick Killen</td>
<td>30</td>
<td>34</td>
</tr>
<tr>
<td>Paul Clough</td>
<td>40</td>
<td>35</td>
</tr>
<tr>
<td>Stephan Johnson</td>
<td>40</td>
<td>37</td>
</tr>
<tr>
<td>Thomas Baker</td>
<td>30</td>
<td>37</td>
</tr>
</tbody>
</table>

data has a positive mutual dependency on its source and target: a worker is more credible if he or she gives a label representing an agreeing edge to a pair of Web pages that are related to the same real world entity; similarly, a worker is less credible if he or she gives a label representing an agreeing edge to a pair of Web pages that are not related to the same real world entity. A disagreeing edge acts as a negative mutual dependency: a worker is less credible if he or she gives a label representing a disagreeing edge to a pair of Web pages that are related to a real world entity; a worker is more credible if he or she gives a label representing a disagreeing edge to a pair of Web pages that are not related to a real world entity.

Given bipartite graph \( G = (T \cup P, E, S) \), where \( T \) is the set of crowd workers, \( P = \{p_1, ..., p_m\} \) is a set of Web page pairs, \( E \subseteq T \times P \) is a set of directed edges derived from the labels given by the workers to pairs of pages, and \( S = \{s_{ij}\} \) is the set of labels attached to the edges. If edge \((t_i, p_j)\) is an agreeing edge, \(s_{ij} = 0\); otherwise, \(s_{ij} = 1\).

Let \( t_i \) be the weight assigned to worker \( t_i \) and \( p_j \) be the weight assigned to Web page pair \( p_j \). The corresponding edge is \((t_i, p_j)\). The use of the mutual dependency principles means that, if \(s_{ij} = 0\), \(t_i\) mirrors \(p_j\), i.e., \(t_i = p_j\); if \(s_{ij} = 1\), \(t_i\) is the opposite of \(p_j\), i.e., \(t_i = 1 - p_j\). The edges are treated equally due to the absence of information on the relative credibility of the connected crowd workers.

\[
\begin{align*}
    t_i &= AVG_{(t_i, p_j) \in E}(1 - 2s_{ij})p_j + s_{ij} \\
    p_j &= AVG_{(t_i, p_j) \in E}(1 - 2s_{ij})t_i + s_{ij}
\end{align*}
\]

These formulae are converted into matrix form, and an iterative algorithm is used to compute the scores for the crowd workers. The score for each worker ranges from 0 to 1. Scores in the range \([0, 0.5]\) are considered workers in low credibility, and those in the range \([0.5, 1]\) are considered high credibility.

### 3. CROWDSOURCED LABELING FOR WEB ENTITY RESOLUTION

In this section, we introduce the method we used for obtaining crowdsourced labels and the majority voting (MV) method we used for aggregating the labels. Then, we evaluate the label quality of an actual crowdsourced data set.

#### 3.1 Data Set

We used data obtained from the Searching Information about Entities in the Web (WEPS) dataset\(^1\), which provides Web pages for given person names, especially for very common names that have high ambiguity on the Web and for names of famous or historical people. This dataset also provides a gold standard for Web pages containing each person’s name, and these Web pages were well clustered by experts. The experts were asked to decide, on the basis of the contents, whether the person names on two Web pages referred to the same person in the real world. The number of Web pages used as training data and as test data for each person name are shown in Table 1 (blank pages were removed). We used tf-idf (the term frequency-inverse document frequency) as the feature.

#### 3.1.1 Crowd-generated Labels

We used the Lancers crowdsourcing service\(^2\) to allocate HITs to crowd workers. Each HIT contained \(n\) Web page URLs. The workers were instructed to open these pages one by one and find all the ones that referred to the same entity in the real world. A drop-down list on the HIT page enabled a worker to assign the same id number to the Web pages considered to refer to the same real world entity. Since we structured the HITs as the clustering of \(N\) Web pages, we had to specify a sampling scheme. If we had structured the HITs as simply separate Web pages in disjoint sets, it would have been impossible to infer a clustering over the entire data set because we would not have known whether two Web pages in different HITs were in the same cluster or not. Therefore, the Web pages had to be members of multiple HITs.

We used a random sampling scheme [19; 9] in which a simple parameter \(V\) is used to control the level of sampling redundancy. We set \(V\) to the number of HITs to which a Web page belonged. This meant that each Web page was voted on \(V\) times by the crowd workers. We set \(n\) to 10 or 20 depending on the number of Web pages available for the person name and \(V\) to 5. After removing the blank Web pages in WEPS, we had 30 Web pages as training data for person names “Armando Valencia,” “Paul Clough,” “Thomas Baker (\(n = 10\)),” 60 for “Abby Watkins,” and 40 for the other person names (\(n=20\)). The total number of data items was \(N\). Thus, the total number of different HITs was

\[
H = NV/n.
\]

Because individual crowd workers often exhibit highly variable labeling accuracy, we asked multiple crowd workers to perform each HIT so that we could derive a single consensus label. Parameter \(R\) is the number of crowd workers who performed the same HIT. Therefore, the total number of HITs generated was

\[
H = NVR/n.
\]

In the experiment described below, each HIT was assumed to be performed five times (\(R = 5\)).

#### 3.1.2 Consensus Labels

Majority voting was used to generate consensus labels for the LPP method in two ways: to derive a single consensus label for each crowd worker and to generate the final consensus labels for all crowd workers. Since Web pages were assigned to multiple HITs, a worker could label a pair of Web pages several times. The two Web pages were assumed to be linked once the same id number was assigned to them by the same person.

As mentioned, simple MV was used to compute the final consensus labels (weighting each worker’s vote equally). Consider a set of noisy labels \(\{y_{ij}\}\) for data items pair \((x_i, x_j)\), where \(y_{ij} \in \{0, 1\}\) is a noisy label provided by the \(t\)-th worker \((t \in \{1, \ldots , T\})\). The set of workers who labeled this pair is defined as \(I_t \subseteq \{1, \ldots , T\}\).
The consensus label \( y_{ij} \) for this pair derived from majority voting is given by

\[
y_{ij} = \begin{cases} 
1 & \text{if } \sum_{t \in J_i} y_{tij} > |J_i|/2, \\
0 & \text{if } \sum_{t \in J_i} y_{tij} < |J_i|/2, \\
\text{random} & \text{otherwise.}
\end{cases}
\]

### 3.2 Label Quality

Using the methods described above, we collected actual crowdsourced data for ten person names ("Abby Watkins, "Alvin Cooper, "Armando Valencia, "David Lodge, "Jerry Hobbs, "Michael Howard, "Patrick Killen, "Paul Clough, "Stephan Johnson, "Thomas Baker") and generated consensus labels. We estimated the label quality of the crowdsourced data by comparing both the labels provided by individual workers and the consensus labels estimated by MV with the gold standard.

The experimental results presented in Figure 2 and Figure 3 and 4 show that the accuracy of the labels provided by individual workers was satisfactory although there was some dispersion of accuracy across workers (the average accuracy was above 80% except for Abby Watkins, and the standard deviation was mostly around 10%). The accuracy for the consensus labels was mostly around 80% and went as high as 90%. Since entity resolution is difficult, the 80% accuracy is higher than predicted and indicates that it is
possible to derive an accurate entity resolution model using these methods.

Here, we give only a general idea of how we can generate consensus labels by using crowdsourcing to resolve entities. Though there are some well-known techniques (e.g., expectation maximization) that may outperform MV, the labeling accuracy of MV is sufficient for our purposes here. The prediction accuracy of the LPP method with consensus labels was less than that of the gold standard no matter how much the labeling accuracy was improved, while experiment results given in section 4 show that the proposed multi-Lap method has accuracy comparable to or better than that of the LPP method with the gold standard.

4. EVALUATION

We experimentally evaluated the prediction accuracy of the LPP method using consensus labels and the multi-Lap method with three weighting measures by implementing it on the Matlab platform (Matlab R2011b) and using the built-in eigs function to solve the generalized eigenvalue problem. The dimension of the latent feature space was set to ten.

We estimated the prediction accuracy of the conventional LPP method [10] for the ground truth labels and for the consensus labels of the crowdsourced data. We also evaluated the prediction accuracy of the multi-Lap method without weighting (Uni), with weighting based on similarity (Sim), and with weighting based on mutual dependency (Mut). We used the AUC (area under the curve) values to measure the prediction.

Generally, for our proposed multi-Lap method, the accuracies using crowd-sourced data were comparable to or better than the conventional LPP method using gold standard and were superior to the LPP method using consensus labels eight out of ten cases in Figure 5.

However, interestingly, the multi-Lap method sometimes outperformed the LPP method using the gold standard, e.g., "Alvin Cooper" (Mut), "Armando Valencia" (Mut). In entity resolution, some Web pages are ambiguous even for human experts. Especially in difficult cases, the labels have to be decided to be a certain value that may cause overfitting. In the multi-Lap method, such difficult cases are treated without aggregating the labels.

By using the mutual dependency principles to calculate the credibility of crowd workers, we can get the correct label even if the majority of the crowd workers give an incorrect label. The multi-Lap method with mutual dependency had better performance than the others two weighting measures five out of ten times and had performance comparable to the LPP method using the gold standard even for two exceptional cases "Patrick Killen" and "Paul Clough".

Comparing the accuracy distributions in Figure 2 with the prediction accuracy, when most workers had virtually the same level of labeling skill ("Jerry Hobbs", "Stephan Johnson"), the LPP method with consensus labels had higher prediction accuracy than the multi-Lap method. However, when most workers were in low labeling skill, as they did for "Abby Watkins", "Patrick Killen" and "Thomas Baker", the multi-Lap method still provided performance comparable to the LPP method using the gold standard.

Furthermore, from the results in Figure 6, we can see that our Multi-lap method always performs as least as fast as the LPP method with the true labels and the LPP method with the consensus labels. Though our Multi-lap with mutual dependency relatively the most time-consuming, most of the time is less than 0.1 second difference.

5. CONCLUSION

We have evaluated the label quality of actual crowdsourced data for entity resolution and compared the prediction accuracy of our proposed combination of multiple Laplacians (with and without weighting) with that of the conventional locality preserving projection method using consensus labels. Analysis of label quality showed that the accuracy of labels provided by individual workers was satisfactory while there was some dispersion of accuracy across workers. The prediction accuracy of the combination of multiple Laplacians method using crowd-sourced labels was comparable to or better than the LPP method using the gold standard and was superior to the LPP method using the consensus labels.

Our combination of multiple Laplacians method with mutual dependency had better performance than the other two weighting measures. However, the multiple Laplacians methods with mutual dependency cannot win in every case, a better weighting measurement is still in need. Our future work is solving the measure used for assigning weights to workers in the learning process.
<table>
<thead>
<tr>
<th>LPP (True Label)</th>
<th>LPP (Consensus Label)</th>
<th>Multi-LAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abby Watkins</td>
<td>0.730</td>
<td>0.654</td>
</tr>
<tr>
<td>Alvin Cooper</td>
<td>0.596</td>
<td>0.502</td>
</tr>
<tr>
<td>Armando Valencia</td>
<td>0.532</td>
<td>0.556</td>
</tr>
<tr>
<td>David Lodge</td>
<td>0.543</td>
<td>0.506</td>
</tr>
<tr>
<td>Jerry Hobbs</td>
<td>0.523</td>
<td>0.560</td>
</tr>
<tr>
<td>Michael Howard</td>
<td>0.545</td>
<td>0.549</td>
</tr>
<tr>
<td>Patrick Killen</td>
<td>0.632</td>
<td>0.538</td>
</tr>
<tr>
<td>Paul Clough</td>
<td>0.528</td>
<td>0.512</td>
</tr>
<tr>
<td>Stephan Johnson</td>
<td>0.657</td>
<td>0.658</td>
</tr>
<tr>
<td>Thomas Baker</td>
<td>0.636</td>
<td>0.533</td>
</tr>
</tbody>
</table>

Figure 5: AUC values.

![Time Cost Graphs](image)

Figure 6: Time cost (CPU: Intel Core i7-3770 3.40GHz*2, memory: 8.00GB, operating system: Windows 7 Professional).

6. REFERENCES
Learning Theory (COLT), 2009.


