Collecting High Quality Overlapping Labels at Low Cost

Grace Hui Yang  
Language Technologies Institute  
Carnegie Mellon University

Anton Mityagin  
Krysta Svore  
Sergey Markov  
Microsoft Bing/Microsoft Research

Roadmap

- Introduction
- How to Use Overlapping Labels
- Selective Overlapping Labeling
- Experiments
- Conclusions and Discussion
Introduction

- Web Search/Learning to Rank
  - Web documents/urls are represented by feature vectors
  - A ranker learns a model from the training data, and computes a rank order of the urls for each query.

- The Web Search Goal
  - Retrieve relevant documents
  - Achieve high retrieval accuracy
    - Measured by NDCG, MAP, or other IR measure

Factors Affecting Retrieval Accuracy

- Amount of training examples
  - The more training examples, the better the accuracy of the trained model
  - Often, large number of training examples are used

- Quality of training labels
  - The higher the quality of labels, the better the accuracy of the trained model
  - How to collect high quality labels?
Affects of Training Data Quality

Solution: Improve *Quality* of Labels

- Label quality depends on
  - Expertise of the labelers
  - The number of labelers
- The more expert the labelers, and the more labelers, the higher the label quality.
- Cost!!!
  - Labelers are expensive
  - High-quality labels can be even more expensive

Figure 1: Learning curves under different quality levels of training data ($q$ is the probability of a label being correct).

Figure cited from [Sheng et al.] KDD’08
Current Approaches

- A lot of (cheap) non-experts for a sample
  - Labelers from Amazon Mechanical Turk
  - Weakness: labels are often unreliable
- Just one label from an expert for a sample
  - The single labeling scheme
  - Widely used in supervised learning
  - Weakness: personal bias
Our Proposed Labeling Scheme

- High quality labels
  - Labels that yield high retrieval accuracy
  - Overlapping labels from experts
- At low cost
  - Only request additional labels when they are needed

Roadmap

- Introduction
- How to Use Overlapping Labels
- Selective Overlapping Labeling
- Experiments
- Conclusion and Discussion
Relevance Labels

- Labels indicate the relevance of a url to a query
  - Perfect, Excellent, Good, Fair, and Bad.

How to Use Overlapping Labels

- How to aggregate overlapping labels?
  - Majority, median, mean, something else?
- Change the weights of the labels?
  - Perfect x 3, Excellent x 2, Good x 2, Bad x 0.5?
- Use overlapping labels only on selected samples?
- How much overlap?
  - 2x, 3x, 5x, 100x?
Aggregating Overlapping Labels

$n$ training samples, $k$ labelers

- **K-Overlap (Using all labels)**
  - When $k=1$, single labeling scheme, training cost: $n$; Labeling cost: 1.
  - Training cost: $kn$; Labeling cost: $k$.

- **Majority vote**
  - Training cost: $n$; Labeling cost: $k$.

- **Highest label**
  - Sort $k$ labels into the order of most-relevant to lease-relevant ($P/E/G/F/B$); Pick the label at the top of the sorted list.
  - Training cost: $n$; Labeling cost: $k$.

Weighting the Labels

- **Assign different weights for labels**
  - Samples labeled as $P/E/G$, assign $w_1$;
  - Samples labeled as $F/B$, assign $w_2$;
  - $w_1 = \theta w_2$, $\theta > 1$.

- **Intuition: “Perfect” probably deserves more weight than other labels**
  - “Perfect” are rare in training data
  - Web search emphasizes on precision

- **Training cost = $n$, Labeling cost = 1.**
Selecting Samples to Label with Overlap

- Collect overlapping labels when it is needed for a sample.

The proposed scheme

Roadmap

- Introduction
- How to Use Overlapping Labels
- Selective Overlapping Labeling
- Experiments
- Conclusion and Discussion
Collect Overlapping Labels When Good+

- **Intuition:**
  - People are difficult to satisfy
    - Seldom say “this url is good”
    - Often say “this url is bad”
  - It is even harder for people to agree on some urls are good

- **So:**
  - If someone thinks a url is good, it is worthwhile to verify with others’ opinions
  - If someone thinks a url is bad, we trust him

If-good-\(k\)

- If a label = P/E/G, get another \(k-1\) overlapping labels;
- Otherwise, keep the first label, go to the next query/url.
- **Example:** (if-good-3)
  - Excellent, Good, Fair
  - Bad
  - Good, Good, Perfect
  - Fair
  - Fair
  - ...

- **Training cost = labeling cost =** \(\frac{n}{r+1} + \frac{nr}{r+1}k\)
- \(r\) is Good+:Fair- ratio among the first labelers.
Good-till-bad

- If a label = P/E/G, get another label;
- If this second label = P/E/G, continue to collect one more label;
- Till a label = F/B.
- Example: (Good-till-bad)
  - Excellent, Good, Fair
  - Bad
  - Good, Good, Perfect, Excellent, Good, Bad
  - Fair
  - ...
- Training cost = labeling cost \( \leq \frac{n}{r+1} + \frac{nr}{r+1}k \).
- Note that \( k \) can be large.

Roadmap

- Introduction
- How to Use Overlapping Labels
- Selective Overlapping Labeling
- Experiments
- Conclusion and Discussion
Datasets

- The Clean label set
  - 2,093 queries; 39,267 query/url pairs
  - 11 labels for each query/url pair
  - 120 judges in total
  - Two feature sets: Clean07 and Clean08

- The Clean+ label set
  - 1,000 queries; 49,785 query/url pairs
  - Created to evaluate if-good-k (k<=3)
  - 17,800 additional labels

Evaluation Metrics

- NDCG for a given query at level $L$:
  \[
  NDCG@L = \frac{1}{Z} \sum_{i=1}^{L} \frac{2^{l(i)} - 1}{\log(1 + i)}
  \]

  where $l(i) = \{0, 1, 2, 3, 4\}$, the relevance label at position $i$;
  $L$: the truncation level.

- NDCG@3, also report @1, @2, @5, @10.
Evaluation

- Average 5~10 runs for an experimental setting
- Two Rankers:
  - LambdaRank [Burges et al. NIPS'06]
  - LambdaMart [Wu et al. MSR-TR-2008-109]

Experimental Settings

- 1. Baseline: the single labeling scheme.
- 2. 3-overlap: 3 overlapping labels, train on all of them.
- 3. 11-overlap: 11 overlapping labels, train on all of them.
- 6. If-good-3: If a label = Good+, get another 2 labels; o/w, keep this label.
- 7. If-good-x3: assign Good+ labels 3 times of weights.
## Retrieval Accuracy on Clean08

### LambdaRank

<table>
<thead>
<tr>
<th>Experiment</th>
<th>NDCG@1</th>
<th>NDCG@2</th>
<th>NDCG@3</th>
<th>NDCG@5</th>
<th>NDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ifgood3</td>
<td>45.03%</td>
<td>45.37%</td>
<td>45.99%*</td>
<td>47.53%</td>
<td>50.53%</td>
</tr>
<tr>
<td>highest3</td>
<td>44.87%</td>
<td>45.17%</td>
<td>45.97%*</td>
<td>47.48%</td>
<td>50.43%</td>
</tr>
<tr>
<td>11-overlap</td>
<td>44.93%</td>
<td>45.10%</td>
<td>45.96%*</td>
<td>47.57%</td>
<td>50.58%</td>
</tr>
<tr>
<td>mv11</td>
<td>44.97%</td>
<td>45.20%</td>
<td>45.89%</td>
<td>47.56%</td>
<td>50.58%</td>
</tr>
<tr>
<td>fgoodx3</td>
<td>44.73%</td>
<td>45.18%</td>
<td>45.80%</td>
<td>47.40%</td>
<td>50.13%</td>
</tr>
<tr>
<td>3-overlap</td>
<td>44.77%</td>
<td>45.27%</td>
<td>45.78%</td>
<td>47.54%</td>
<td>50.50%</td>
</tr>
<tr>
<td>mv3</td>
<td>44.83%</td>
<td>45.11%</td>
<td>45.66%</td>
<td>47.09%</td>
<td>49.83%</td>
</tr>
<tr>
<td>goodtilbad</td>
<td>44.88%</td>
<td>44.87%</td>
<td>45.58%</td>
<td>47.05%</td>
<td>49.86%</td>
</tr>
<tr>
<td>baseline</td>
<td>44.72%</td>
<td>44.98%</td>
<td>45.53%</td>
<td>46.93%</td>
<td>49.69%</td>
</tr>
</tbody>
</table>

Gain on Clean08 (LambdaRank): 0.46 point NDCG@3

### LambdaMart

<table>
<thead>
<tr>
<th>Experiment</th>
<th>NDCG@1</th>
<th>NDCG@2</th>
<th>NDCG@3</th>
<th>NDCG@5</th>
<th>NDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ifgood3</td>
<td>44.63%</td>
<td>45.08%</td>
<td>45.93%*</td>
<td>47.65%</td>
<td>50.37%</td>
</tr>
<tr>
<td>11-overlap</td>
<td>44.70%</td>
<td>45.13%</td>
<td>45.91%*</td>
<td>47.59%</td>
<td>50.35%</td>
</tr>
<tr>
<td>mv11</td>
<td>44.31%</td>
<td>44.86%</td>
<td>45.48%</td>
<td>47.02%</td>
<td>49.97%</td>
</tr>
<tr>
<td>highest3</td>
<td>44.46%</td>
<td>44.81%</td>
<td>45.42%</td>
<td>47.16%</td>
<td>50.09%</td>
</tr>
<tr>
<td>fgoodx3</td>
<td>43.78%</td>
<td>44.14%</td>
<td>44.80%</td>
<td>46.42%</td>
<td>49.26%</td>
</tr>
<tr>
<td>3-overlap</td>
<td>43.52%</td>
<td>44.23%</td>
<td>44.77%</td>
<td>46.49%</td>
<td>49.44%</td>
</tr>
<tr>
<td>baseline</td>
<td>43.48%</td>
<td>43.89%</td>
<td>44.45%</td>
<td>46.11%</td>
<td>49.12%</td>
</tr>
<tr>
<td>mv3</td>
<td>42.96%</td>
<td>43.25%</td>
<td>44.01%</td>
<td>45.56%</td>
<td>48.30%</td>
</tr>
</tbody>
</table>

Gain on Clean08 (LambdaMart): 1.48 point NDCG@3
Retrieval Accuracy on Clean+
LambdaRank

<table>
<thead>
<tr>
<th>Experiment</th>
<th>NDCG@1</th>
<th>NDCG@2</th>
<th>NDCG@3</th>
<th>NDCG@5</th>
<th>NDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ifgood2</td>
<td>50.53%</td>
<td>49.03%</td>
<td>48.57%**</td>
<td>48.56%</td>
<td>50.02%</td>
</tr>
<tr>
<td>ifgood3</td>
<td>50.33%</td>
<td>48.84%</td>
<td>48.41%</td>
<td>48.48%</td>
<td>49.89%</td>
</tr>
<tr>
<td>baseline</td>
<td>50.32%</td>
<td>48.72%</td>
<td>48.20%</td>
<td>48.31%</td>
<td>49.65%</td>
</tr>
<tr>
<td>ifgoodx3</td>
<td>50.04%</td>
<td>48.51%</td>
<td>48.16%</td>
<td>48.18%</td>
<td>49.61%</td>
</tr>
</tbody>
</table>

Gain on Clean+: 0.37 point NDCG@3

NDCG@3 for If-Good-k Runs (Clean08,
LambdaRank)
## Costs of Overlapping Labeling

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Labeling Cost</th>
<th>Training Cost</th>
<th>Fair-: Good+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1</td>
<td>1</td>
<td>3.72</td>
</tr>
<tr>
<td>3-overlap</td>
<td>3</td>
<td>3</td>
<td>3.71</td>
</tr>
<tr>
<td>mv3</td>
<td>3</td>
<td>1</td>
<td>4.49</td>
</tr>
<tr>
<td>mv11</td>
<td>11</td>
<td>1</td>
<td>4.37</td>
</tr>
<tr>
<td>If-good-3</td>
<td>1.41</td>
<td>1.41</td>
<td>2.24</td>
</tr>
<tr>
<td>If-good-x3</td>
<td>1</td>
<td>1.41</td>
<td>2.24</td>
</tr>
<tr>
<td>Highest-3</td>
<td>3</td>
<td>1</td>
<td>1.78</td>
</tr>
<tr>
<td>Good-till-bad</td>
<td>1.87</td>
<td>1.87</td>
<td>1.38</td>
</tr>
<tr>
<td>11-overlap</td>
<td>11</td>
<td>11</td>
<td>4.37</td>
</tr>
</tbody>
</table>

## Discussion

- Why if-good-2/3 works?
  - More balanced training dataset?
  - More positive training samples?
  - No! (since simple weighting does not perform well)
Discussion

Why if-good-2/3 works?
- Better capture the worthiness of reconfirming a judgment
- Yield higher quality labels

Discussion

Why does it only need 1 or 2 additional labels?
- Too many opinions from different labelers may create too much noise and too high variance.
Conclusions

- If-good-k is statistically better than single labeling; and statistically better than other methods in most cases
- Only 1 or 2 additional labels are needed for selected sample
- If-good-2/3 is cheap in labeling cost: ~1.4.
- What doesn’t work:
  - Majority vote
  - Simply change weights for labels

Thanks and Questions?

- Contact:
  - huiyang@cs.cmu.edu
  - mityagin@gmail.com
  - ksvore@microsoft.com
  - sergey.markov@microsoft.com
Retrieval Accuracy on Clean07

### LambdaRank

<table>
<thead>
<tr>
<th>Experiment</th>
<th>NDCG@1</th>
<th>NDCG@2</th>
<th>NDCG@3</th>
<th>NDCG@5</th>
<th>NDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ifgood3</td>
<td>46.23%</td>
<td>47.80%</td>
<td>49.55**</td>
<td>51.81%</td>
<td>55.38%</td>
</tr>
<tr>
<td>mv11</td>
<td>45.77%</td>
<td>47.76%</td>
<td>49.30%</td>
<td>51.60%</td>
<td>55.09%</td>
</tr>
<tr>
<td>goodtillbad</td>
<td>45.72%</td>
<td>47.80%</td>
<td>49.22%</td>
<td>51.73%</td>
<td>55.23%</td>
</tr>
<tr>
<td>highest3</td>
<td>45.75%</td>
<td>47.67%</td>
<td>49.16%</td>
<td>51.49%</td>
<td>55.01%</td>
</tr>
<tr>
<td>3-overlap</td>
<td>45.52%</td>
<td>47.48%</td>
<td>49.00%</td>
<td>51.51%</td>
<td>54.90%</td>
</tr>
<tr>
<td>ifgoodx3</td>
<td>45.25%</td>
<td>47.28%</td>
<td>48.98%</td>
<td>51.26%</td>
<td>54.82%</td>
</tr>
<tr>
<td>mv3</td>
<td>45.07%</td>
<td>47.28%</td>
<td>48.87%</td>
<td>51.36%</td>
<td>54.93%</td>
</tr>
<tr>
<td>11-overlap</td>
<td>45.25%</td>
<td>47.24%</td>
<td>48.69%</td>
<td>51.11%</td>
<td>54.58%</td>
</tr>
<tr>
<td>baseline</td>
<td>45.18%</td>
<td>47.06%</td>
<td>48.60%</td>
<td>51.02%</td>
<td>54.51%</td>
</tr>
</tbody>
</table>

**Gain on Clean07 (LambdaRank): 0.95 point NDCG@3**

### LambdaMart

<table>
<thead>
<tr>
<th>Experiment</th>
<th>NDCG@1</th>
<th>NDCG@2</th>
<th>NDCG@3</th>
<th>NDCG@5</th>
<th>NDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ifgood3</td>
<td>44.63%</td>
<td>45.08%</td>
<td>45.93%*</td>
<td>47.65%</td>
<td>50.37%</td>
</tr>
<tr>
<td>3-overlap</td>
<td>44.70%</td>
<td>45.13%</td>
<td>45.91%*</td>
<td>47.59%</td>
<td>50.35%</td>
</tr>
<tr>
<td>11-overlap</td>
<td>44.31%</td>
<td>44.86%</td>
<td>45.48%</td>
<td>47.02%</td>
<td>49.97%</td>
</tr>
<tr>
<td>mv11</td>
<td>44.46%</td>
<td>44.81%</td>
<td>45.42%</td>
<td>47.16%</td>
<td>50.09%</td>
</tr>
<tr>
<td>ifgoodx3</td>
<td>43.78%</td>
<td>44.14%</td>
<td>44.80%</td>
<td>46.42%</td>
<td>49.26%</td>
</tr>
<tr>
<td>highest3</td>
<td>43.52%</td>
<td>44.23%</td>
<td>44.77%</td>
<td>46.49%</td>
<td>49.44%</td>
</tr>
<tr>
<td>mv3</td>
<td>43.48%</td>
<td>43.89%</td>
<td>44.45%</td>
<td>46.11%</td>
<td>49.12%</td>
</tr>
<tr>
<td>baseline</td>
<td>42.96%</td>
<td>43.25%</td>
<td>44.01%</td>
<td>45.56%</td>
<td>48.30%</td>
</tr>
</tbody>
</table>

**Gain on Clean07 (LambdaMart): 1.92 point NDCG@3**