# 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

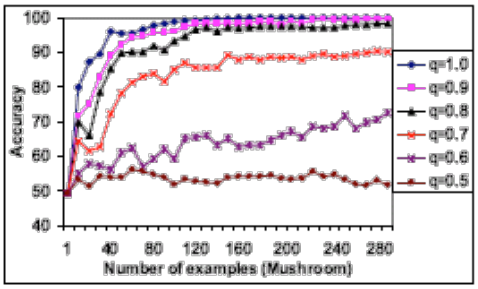



**Figure 1: Learning curves under different quality levels of training data ($q$ is the probability of a label being correct).**

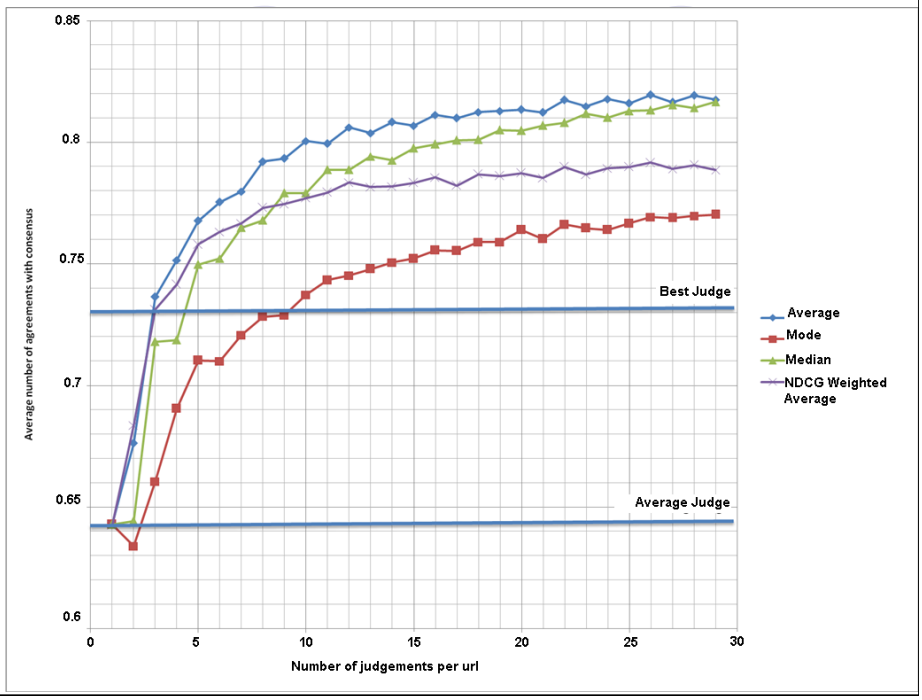
Figure cited from [Sheng et al.] KDD' 08

## 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

# 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?
  - Perfectx3, Excellentx2, Goodx2, Badx0.5?
- 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.


## 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*:

$$NDCG@L = \frac{1}{Z}\sum_{i=1}^{L}\frac{2^{l(i)}-1}{\log(1+i)}$$

$l(i) = \{0,1,2,3,4\}$ , 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.
- 4. Mv3: Majority Vote of 3 labels.
- 5. Mv11: Majority Vote of 11 labels.
- 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.
- 8. Highest-3: The highest label among 3 labels.
- 9. Good-till-bad: k=11.

# Retrieval Accuracy on Clean08 LambdaRank

| Experiment | NDCG@1 | NDCG@2 | NDCG@3 | NDCG@5 | NDCG@10 |
|---|---|---|---|---|---|
| ifgood3 | 45.03% | 45.37% | **45.99%*** | 47.53% | 50.53% |
| highest3 | 44.87% | 45.17% | **45.97%*** | 47.48% | 50.43% |
| 11-overlap | 44.93% | 45.10% | **45.96%*** | 47.57% | 50.58% |
| mv11 | 44.97% | 45.20% | 45.89% | 47.56% | 50.58% |
| ifgoodx3 | 44.73% | 45.18% | 45.80% | 47.40% | 50.13% |
| 3-overlap | 44.77% | 45.27% | 45.78% | 47.54% | 50.50% |
| mv3 | 44.83% | 45.11% | 45.66% | 47.09% | 49.83% |
| goodtillbad | 44.88% | 44.87% | 45.58% | 47.05% | 49.86% |
| baseline | 44.72% | 44.98% | 45.53% | 46.93% | 49.69% |

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

# Retrieval Accuracy on Clean08 LambdaMart

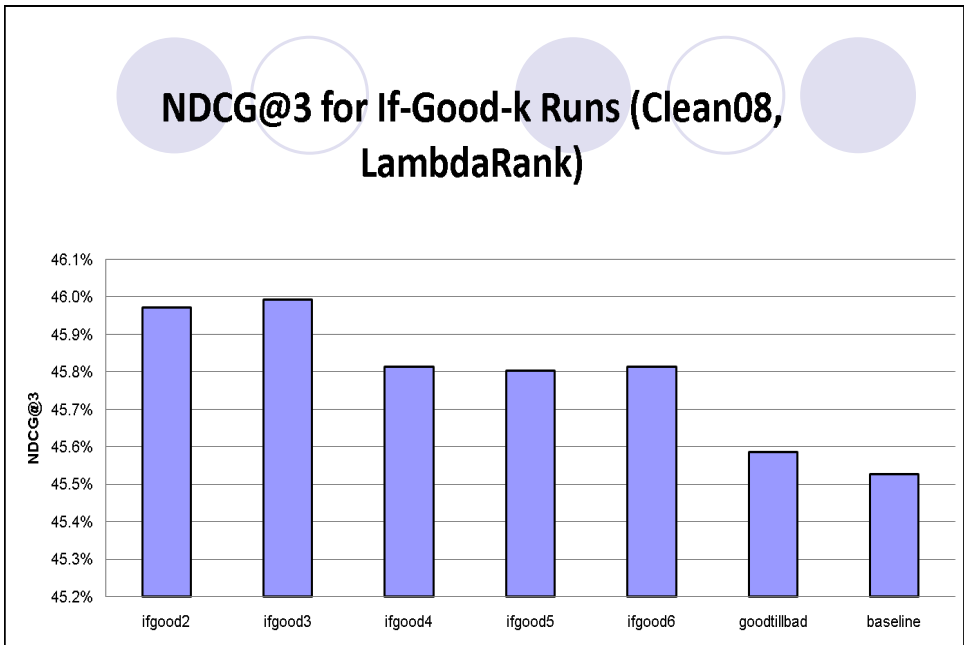
| Experiment | NDCG@1 | NDCG@2 | NDCG@3 | NDCG@5 | NDCG@10 |
|---|---|---|---|---|---|
| ifgood3 | 44.63% | 45.08% | **45.93%*** | 47.65% | 50.37% |
| 11-overlap | 44.70% | 45.13% | **45.91%*** | 47.59% | 50.35% |
| mv11 | 44.31% | 44.86% | 45.48% | 47.02% | 49.97% |
| highest3 | 44.46% | 44.81% | 45.42%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

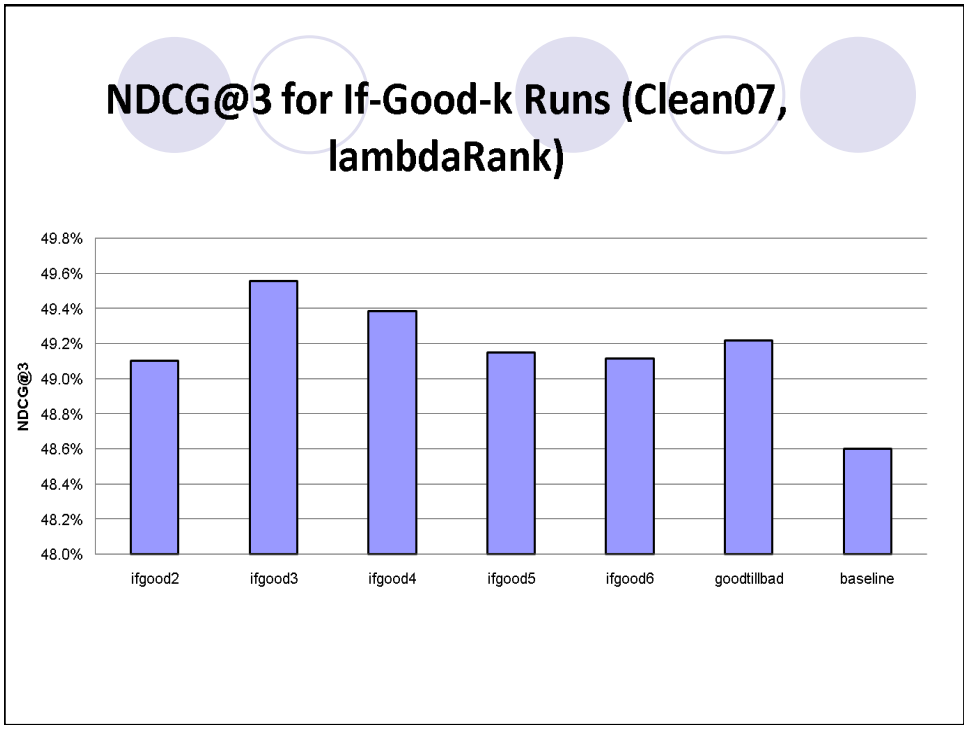
**Gain on Clean08 (LambdaMart): 1.48 point NDCG@3**

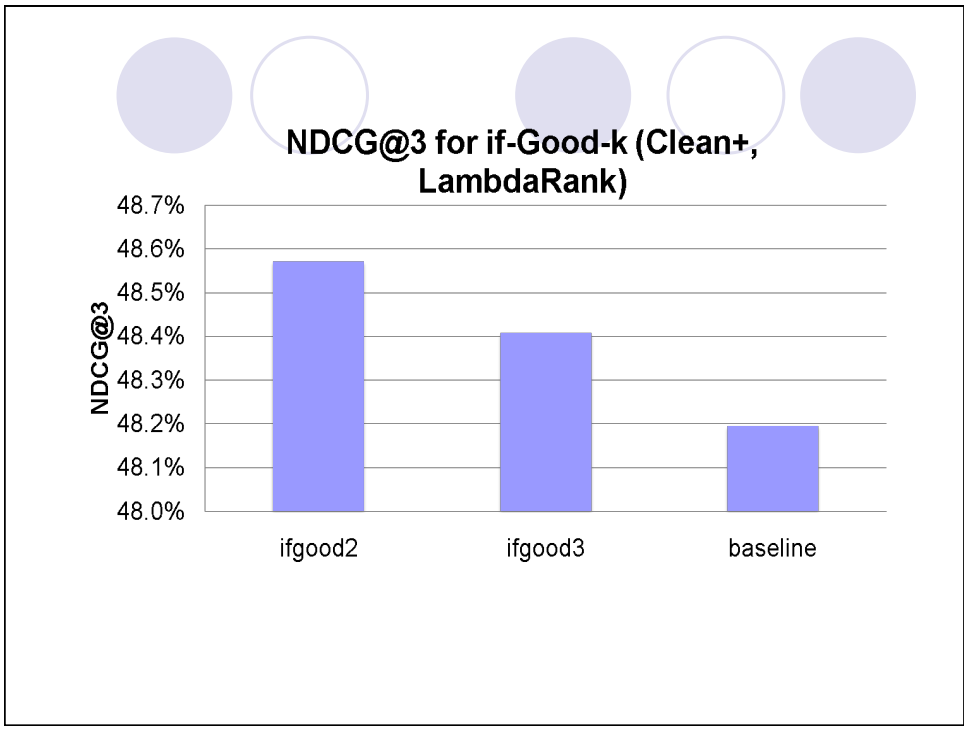
# Retrieval Accuracy on Clean+ LambdaRank

| Experiment | NDCG@1 | NDCG@2 | NDCG@3 | NDCG@5 | NDCG@10 |
|---|---|---|---|---|---|
| ifgood2 | 50.53% | 49.03% | **48.57%**** | 48.56% | 50.02% |
| ifgood3 | 50.33% | 48.84% | 48.41% | 48.48% | 49.89% |
| baseline | 50.32% | 48.72% | 48.20% | 48.31% | 49.65% |
| ifgoodx3 | 50.04% | 48.51% | 48.16% | 48.18% | 49.61% |

**Gain on Clean+: 0.37 point NDCG@3**




NDCG@3 for If-Good-k Runs (Clean07, lambdaRank)
NDCG@3
49.8%
49.6%
49.4%
49.2%
49.0%
48.8%
48.6%
48.4%
48.2%
48.0%
ifgood2
ifgood3
ifgood4
ifgood5
ifgood6
goodtillbad
baseline

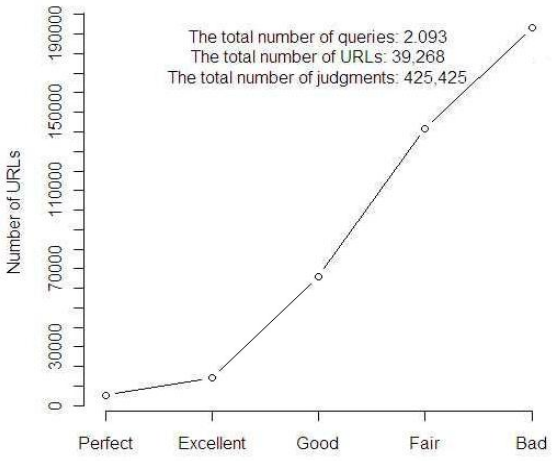

NDCG@3 for if-Good-k (Clean+, LambdaRank)
NDCG@3
48.7%
48.6%
48.5%
48.4%
48.3%
48.2%
48.1%
48.0%
ifgood2
ifgood3
baseline

# Costs of Overlapping Labeling(Clean07)

| Experiment | Labeling Cost | Training Cost | Fair-: Good+ |
|---|---|---|---|
| Baseline | 1 | 1 | 3.72 |
| 3-overlap | 3 | 3 | 3.71 |
| mv3 | 3 | 1 | 4.49 |
| mv11 | 11 | 1 | 4.37 |
| **lf-good-3** | **1.41** | **1.41** | **2.24** |
| lf-good-x3 | 1 | 1.41 | 2.24 |
| Highest-3 | 3 | 1 | 1.78 |
| Good-till-bad | 1.87 | 1.87 | 1.38 |
| 11-overlap | 11 | 11 | 4.37 |

# 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

| Experiment | NDCG@1 | NDCG@2 | NDCG@3 | NDCG@5 | NDCG@10 |
|---|---|---|---|---|---|
| ifgood3 | 46.23% | 47.80% | 49.55%** | 51.81% | 55.38% |
| mv11 | 45.77% | 47.76% | 49.30% | 51.60% | 55.09% |
| goodtillbad | 45.72% | 47.80% | 49.22% | 51.73% | 55.23% |
| Highest3 | 45.75% | 47.67% | 49.16% | 51.49% | 55.01% |
| 3-overlap | 45.52% | 47.48% | 49.00% | 51.51% | 54.90% |
| ifgoodx3 | 45.25% | 47.28% | 48.98% | 51.26% | 54.82% |
| mv3 | 45.07% | 47.28% | 48.87% | 51.36% | 54.93% |
| 11-overlap | 45.25% | 47.24% | 48.69% | 51.11% | 54.58% |
| baseline | 45.18% | 47.06% | 48.60% | 51.02% | 54.51% |

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

# Retrieval Accuracy on Clean07 LambdaMart

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

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