Are Click-through Data Adequate for Learning Web Search Rankings?

Zhicheng Dou, Ruihua Song, Xiaojie Yuan*, and Ji-Rong Wen
Microsoft Research Asia, *Nankai University
2008.10

Outline

- Motivation
- Related works
- Methodology
- Datasets
- Preferences Extraction Strategies
- Correlation between HRS and CT
- Effectiveness of CT for learning to rank
- Conclusions

Motivation

- Learning-to-rank algorithms:
  - Ranking SVM, RankNet, RankBoost, LamdaRank, et al.
- Training data
  - Human Judgments
    - costly and time-consuming
    - Limited relevance levels
    - Difficult, especially for ambiguous queries
  - Clickthrough data
    - Easy to get, unlimited amount
    - Decisions of *a large number* of *real-world* users
- Motivation questions
  - What is the reliability of CT?
  - Are CT useful and effective in learning to rank?

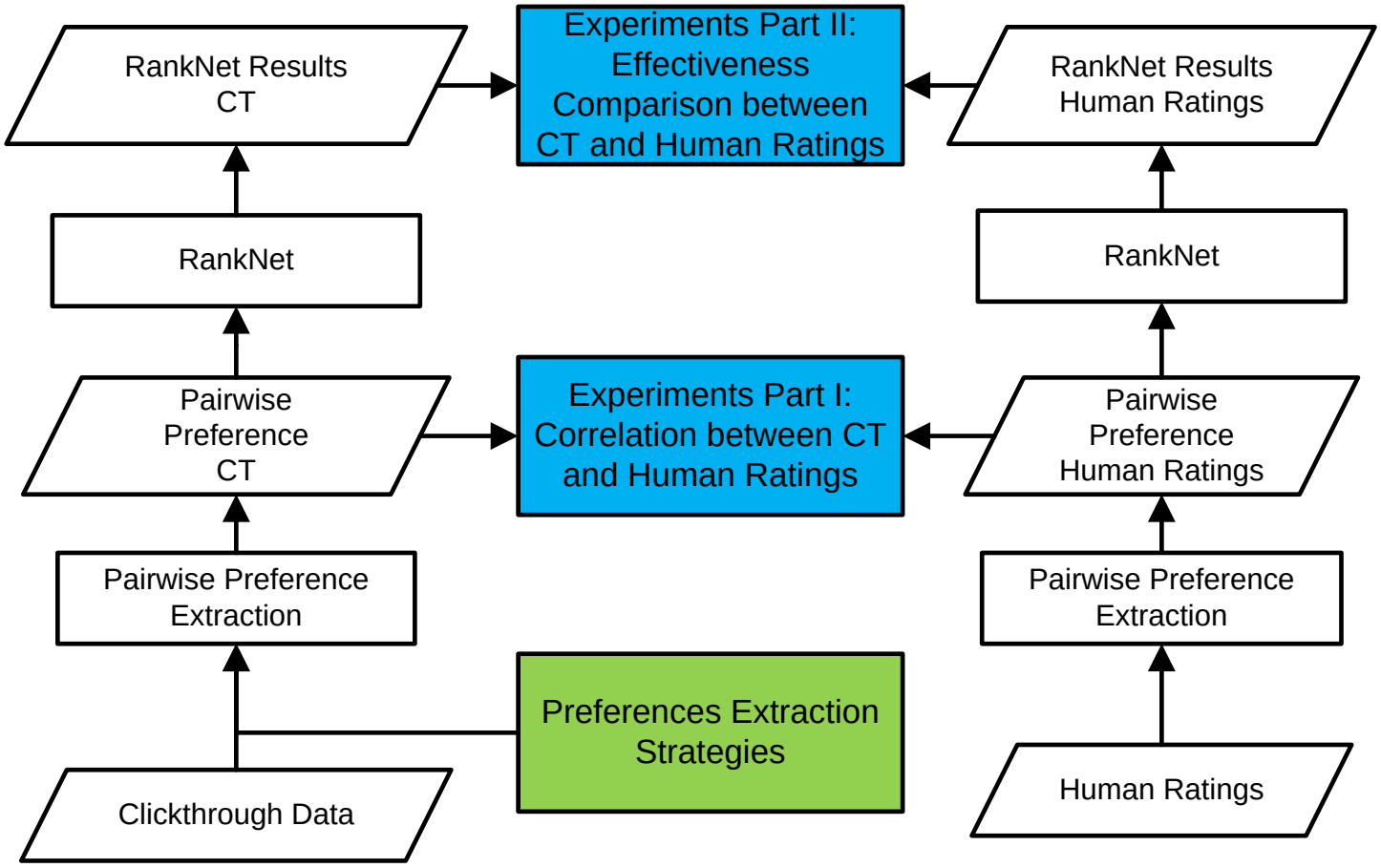
Related works

- Joachims et al.
  - Extract reliable pairs from individual queries and query chains (e.g., click>non-click above)
  - Laboratorial settings
- Agichtein, Brill, and Dumais
  - User behavior is used as features
  - Large amounts of human judgments are still needed

Our approach

- Aggregate user clicks for each query-document pair
- Generate training examples (relative preferences, document pairs) by comparing aggregated click frequencies
- Use preferences to Learn and Evaluate ranking

Framework


RankNet Results
CT
Experiments Part II:
Effectiveness
Comparison between
CT and Human Ratings
RankNet Results
Human Ratings
RankNet
RankNet
Pairwise
Preference
CT
Experiments Part I:
Correlation between CT
and Human Ratings
Pairwise
Preference
Human Ratings
Pairwise Preference
Extraction
Pairwise Preference
Extraction
Preferences Extraction
Strategies
Clickthrough Data
Human Ratings

Datasets

- Human Rating Data (HRS Data)
  - 10,000 training, 1,000 validation, and 1,000 test
- Clickthrough Data
  - 46 days (July 9, 2007 to August 23, 2007)
  - Calculate an aggregated click frequency for each query-document pair
  - Ignore other information, e.g., click position
- Format: Query ID, Doc ID, Rating, *ClickFreq*, {features}

Table 1: Basic statistics of dataset.

	Training	Validation	Test
#Queries	10,000	1,000	1,000
#Documents	584,322	325,514	324,782
#Judged Documents	313,316	28,820	29,788
#Clicked Documents	71,170	6,301	6,880

Preference Extraction Strategies

- Use pair-wise relevance preferences
- Strategies
  - **Label**: If rating(q, d_i) > rating(q, d_j), a relevance preference example $rel\,(q, d_i) >_{lbl} rel(q, d_j)$ is extracted
  - **CT**: Let click frequency difference $cdiff(q, d_i, d_j) = click(q, d_i) - click(q, d_j)$ If $cdiff(q, d_i, d_j) > 0$, a relevance preference example $rel\,(q, d_i) >_{ct} rel(q, d_j)$ is extracted.
  - **CT_Gn**: A relevance preference example $rel\,(q, d_i) >_{ct} rel\,(q, d_j)$ is extracted only when $cdiff(q, d_i, d_j) > n$.

Part I: Correlation between Label and CT

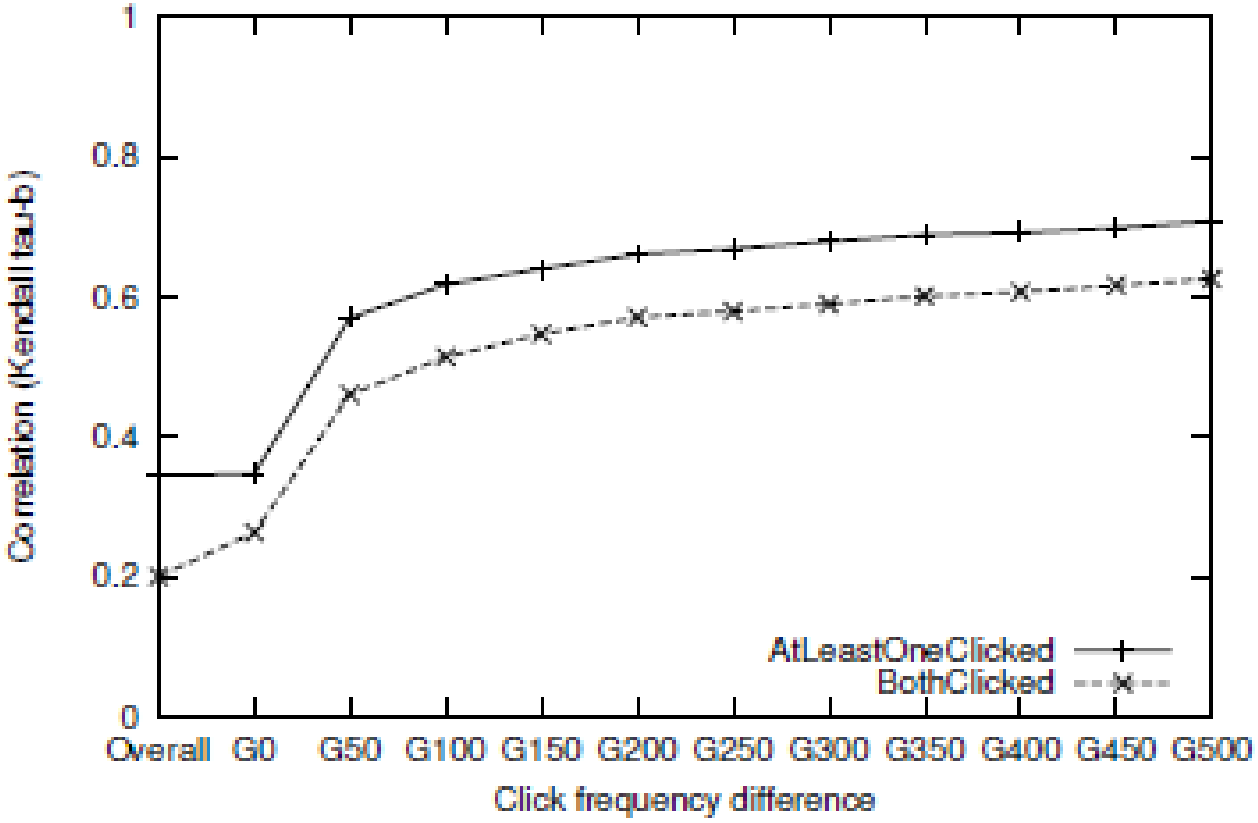
- Using Kendall Tau-b
- Results
  - Small correlation values between CT and human ratings
  - CT correlates more to human judgments when including un-clicked documents(AtLeastOneClicked)

Table 2: Overall correlation between click-through data and human judgments (Kendall tau-b)

	Training	Validation	Test
BothClicked	0.201274	0.163600	0.194758
AtLeastOneClicked	0.345716	0.300375	0.363094

Part I: Correlation between Label and CT

- Click frequency difference (CT_Gn)
  - Pairs with larger click frequency differences correlate more to human judgments



Part I: Correlation between Label and CT

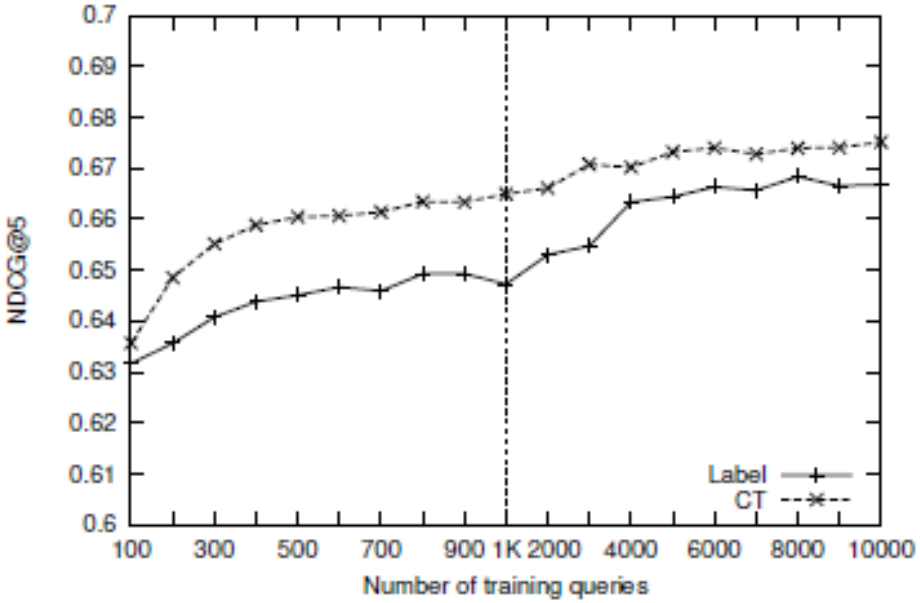
- Summary
  - Clickthrough data and human ratings are not totally same
  - Pairs with larger click frequency differences correlate more to human judgments

Part II: Effectiveness of CT for learning to rank

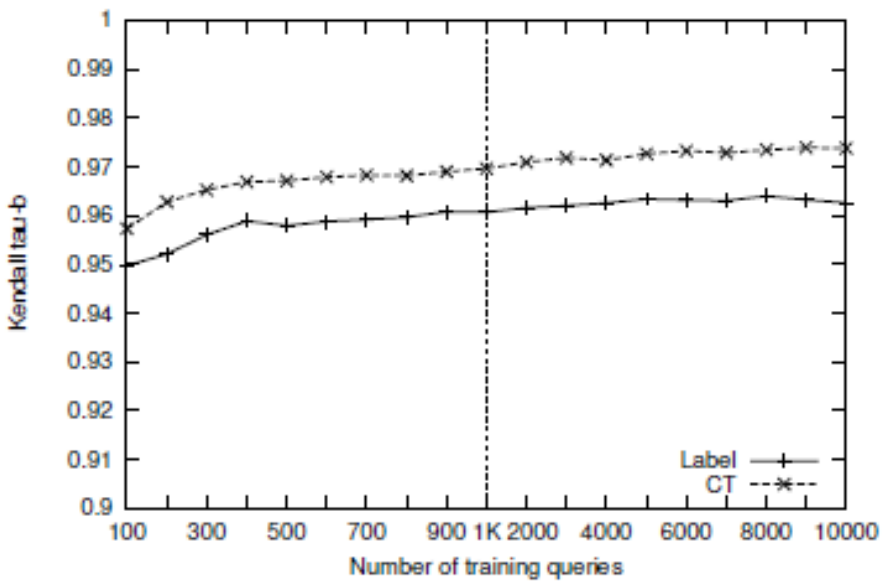
- Use RankNet
- Train RankNet using pairwise preferences
- Evaluation metrics
  - NDCG@5, based upon *human ratings*
  - Kendall Tau-b, based upon clickthrough

Part II: Effectiveness of CT for learning to rank

- Overall results
  - CT outperforms Label with all sizes of training set when using equivalent queries
  - **Preferences in CT are more useful and effective for learning, even using a straightforward preference generation strategy**



(a) NDCG@5



(b) Kendall tau-b

Part II: Effectiveness of CT for learning to rank

- Click frequency Differences (CT_Gn)
  - Pairs with larger click frequency differences do not achieve get better performance
    - Possible reason: Much Less pairs

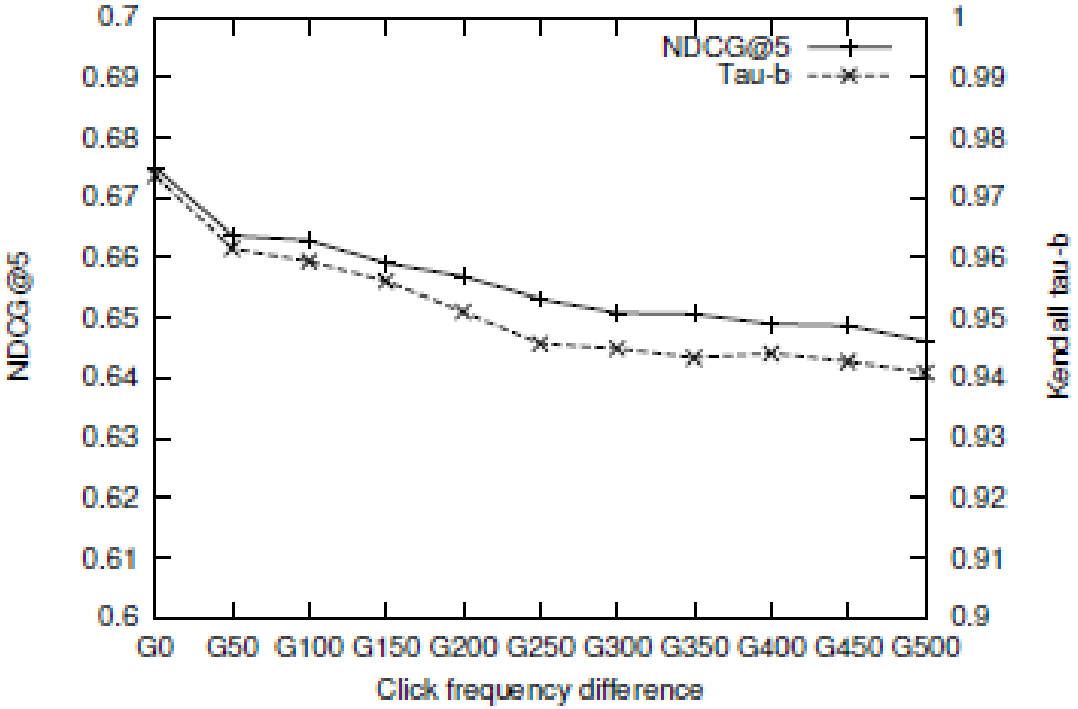


Figure 5: RankNet performance when using pairs with variant click frequency differences

Part II: Effectiveness of CT for learning to rank

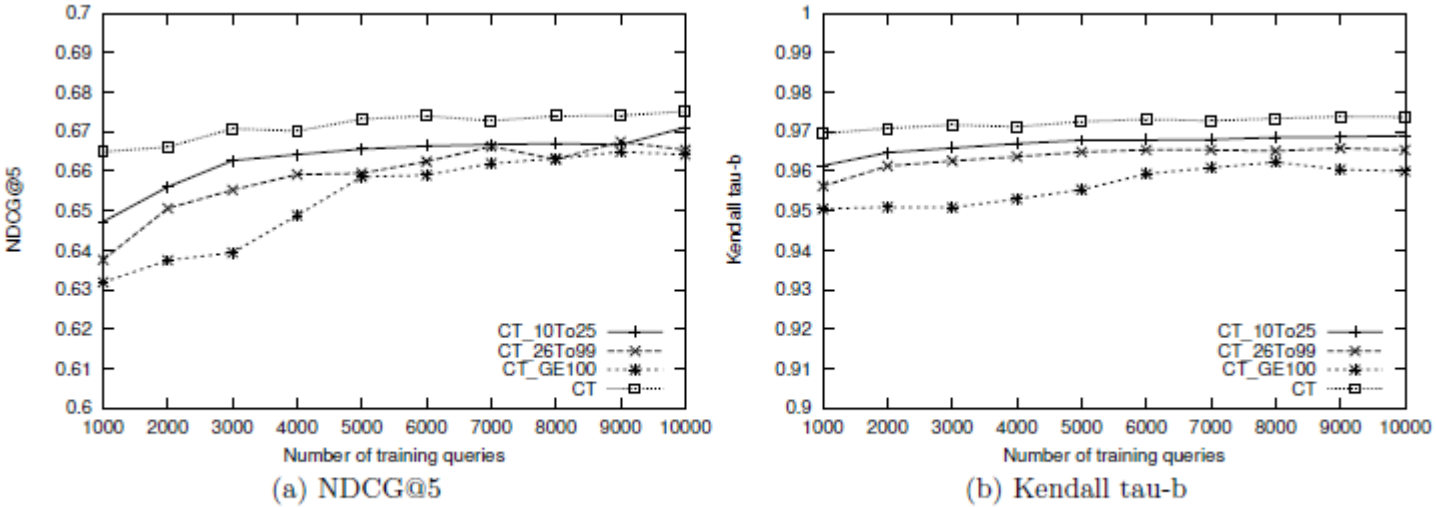
- Three preference selection strategies
  - 10To25, 26To99, and GE100
  - Equal amounts of training examples
- Correlation with human ratings
  - GE100>26To99>10To25

Table 3: Correlation between human judgments and click-through data under three different pair selection strategies

	Training	Validation	Test
CT_10To25	0.305743	0.270589	0.306940
CT_26To99	0.390308	0.361930	0.337831
CT_GE100	0.617736	0.628011	0.605718

Part II: Effectiveness of CT for learning to rank

- Learning performance
  - 10To25>26To99>GE100!!!
- Possible reasons:
  - Pairs with larger click frequency differences
    - More reliable, but simple, biased, contain limited information
  - Pairs with smaller click frequency differences
    - Are more comprehensive and informative

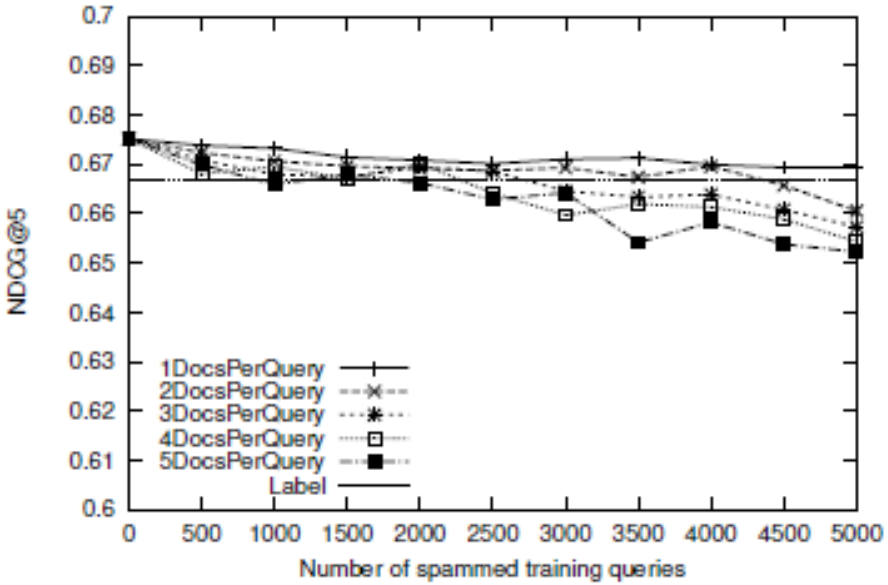


(a) NDCG@5 (b) Kendall tau-b

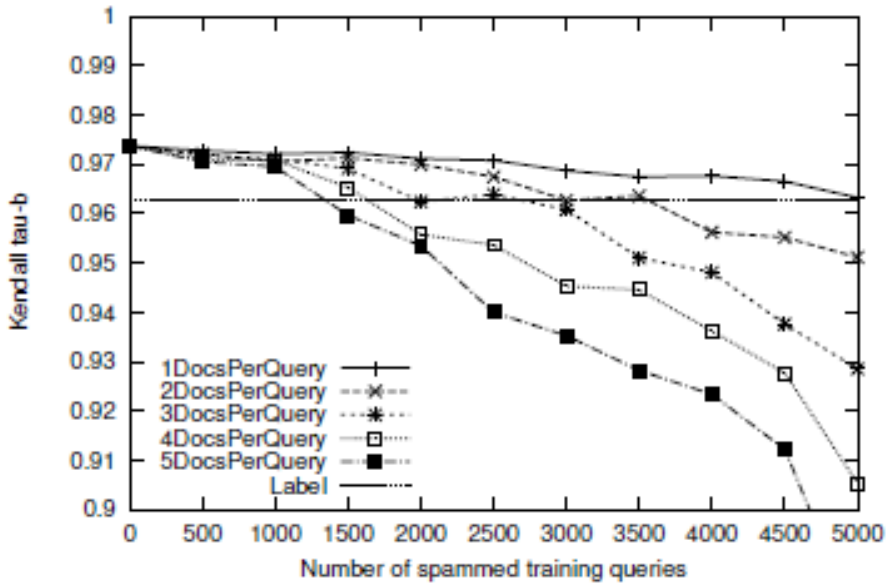
Figure 7: RankNet performance of three different pair selection strategies

Part II: Effectiveness of CT for learning to rank

- Stability
  - CT is less sensitive to click spam



(a) NDCG@5



(b) Kendall tau-b

Figure 9: RankNet performance when variant numbers of queries are spammed.

Key conclusions/contributions

- Conclusions
  - Click-through data are effective for learning web search rankings, even better than human judgments;
  - Click-through data can be more reliable, more comprehensive, and more informative than human judgments in some cases;
  - Reliability and coverage of training data are both important for learning.

Future Work

- Click-through Modeling
  - Position Bias Removing
- Combination of Click-through and Human Judgments

Questions?

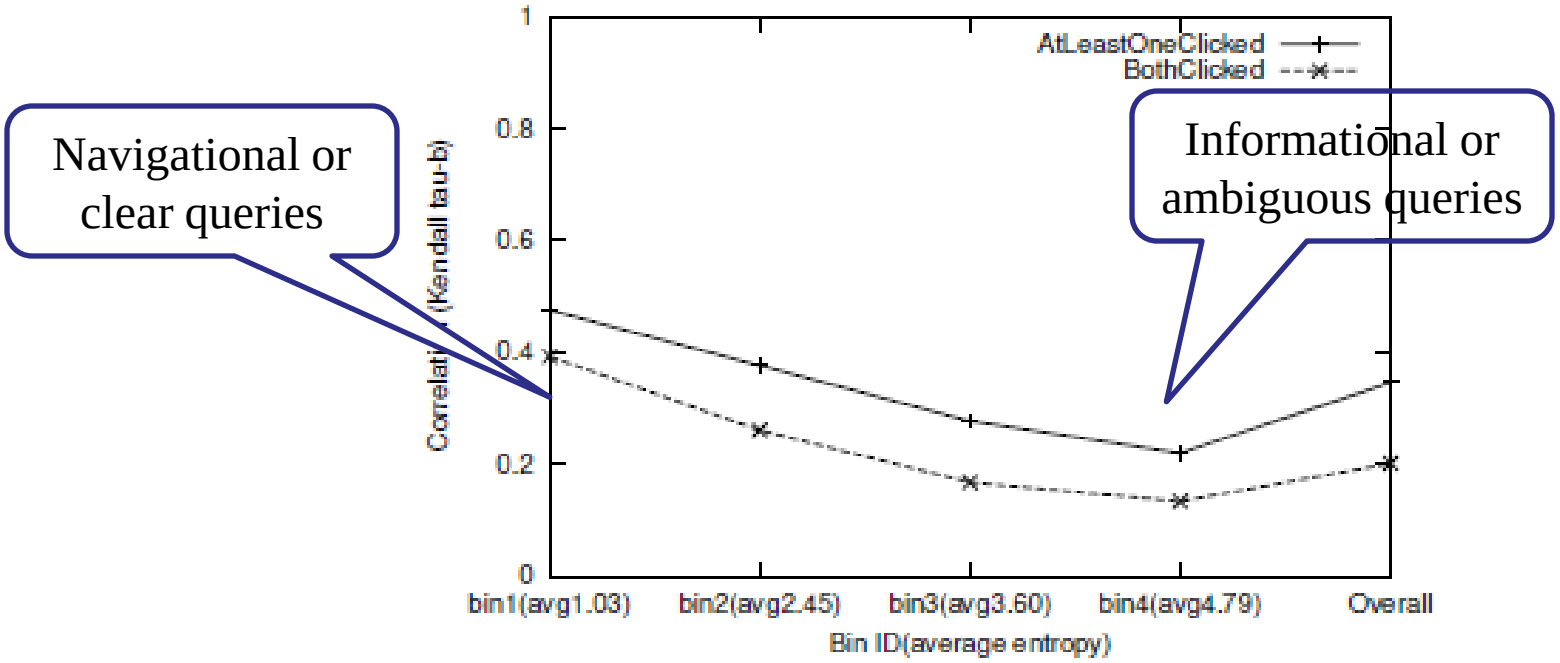
Thanks.

Part I: Correlation between HRS and CT

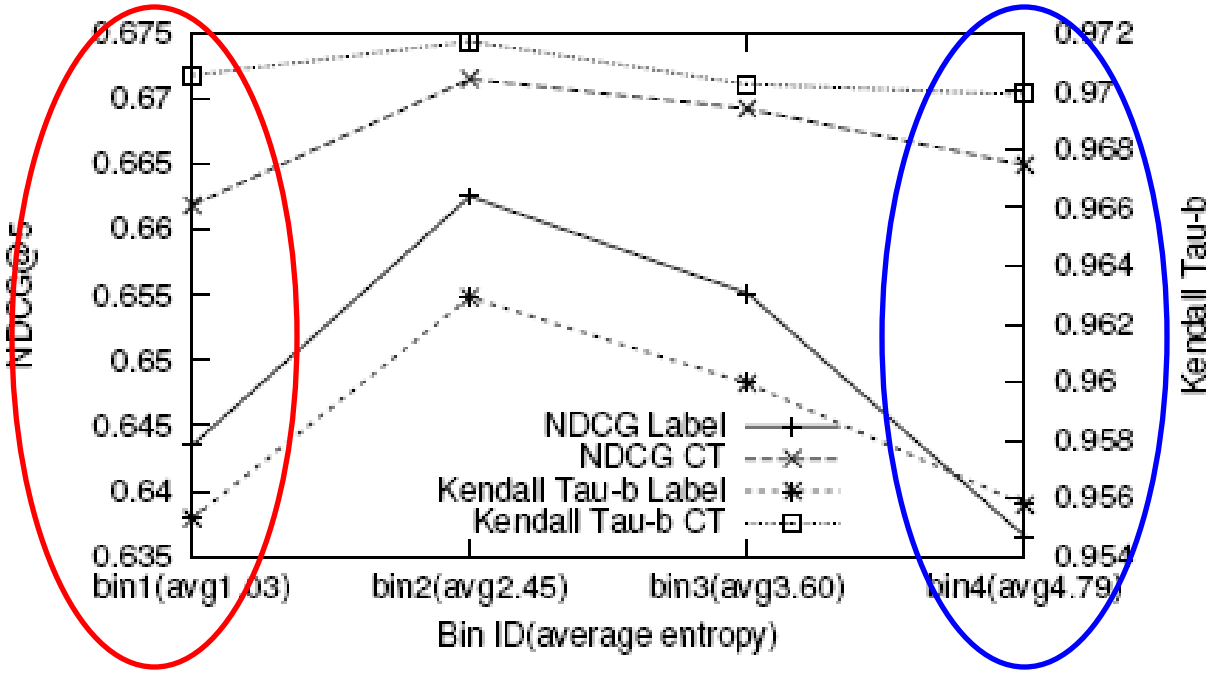
- Click Entropy

$$ClickEntropy(q) = \sum_{d \in D(q)} -P(d \mid q) \log_2 P(d \mid q)$$

  - CT correlates more to human judgments for queries with smaller click entropies



Part II: Effectiveness of CT for learning to rank- **Click Entropy**



- Pairs in human ratings:
  - Bin1 : biased
  - Bin4: Less reliable
- Pairs in CT: more robust for learning
  - Bin1: more comprehensive
  - Bin4: more reliable