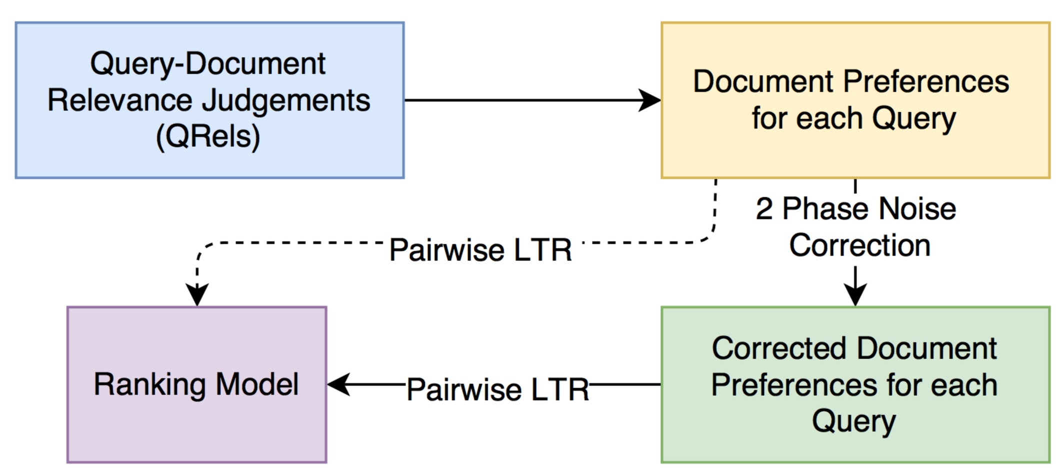
Noise Correction in Pairwise Document Preferences for Learning to Rank

Harsh Trivedi & Prasenjit Majumder
harshjtrivedi94@gmail.com & prasenjit.majumder@gmail.com

1. Motivation

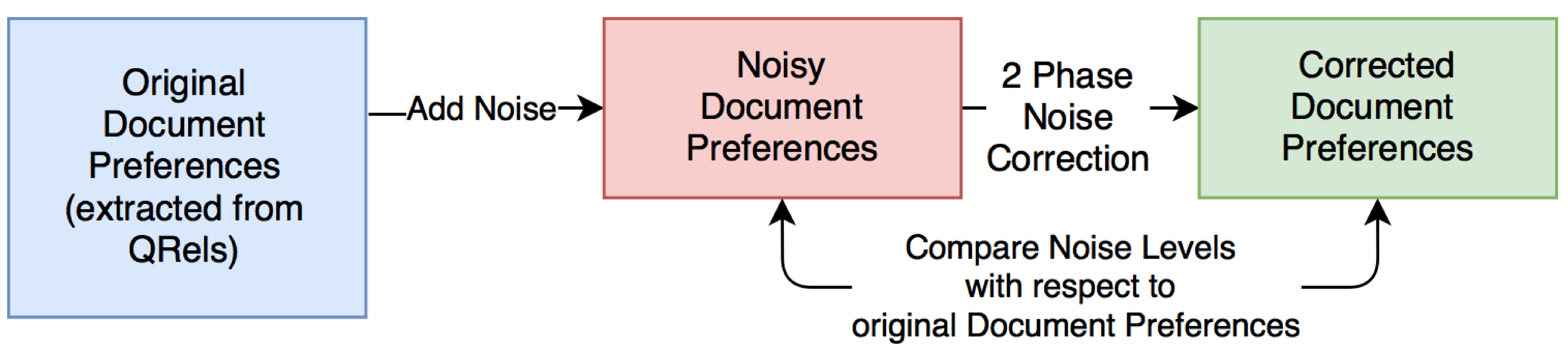
- Learning to Rank (LTR) is the application of supervised machine learning in the construction of ranking models for information retrieval systems.
- Previous works have shown that human judges may not agree with each other in the task of assigning relevance labels to query-document pairs which induces noise in training data [1].
- Studies have shown that noise in training data for Learning to Rank negatively affects the learned ranking model [2].
- Types of Learning to Rank Algorithms: *Pointwise*, *Pairwise*, *Listwise*. Focus on this paper is on the pairwise algorithm which takes pairwise document preferences for each query separately to learn the ranking model. Eg. Rank SVM, Rank Net, GB Rank, QB Rank etc.
- This paper proposes a way of correcting noise in the training data for pairwise LTR algorithms.

2. Overall Objective



- **End Goal:** Show that correcting errors in training document preferences can improve Pairwise LTR performance.
- In this short paper, building upon [2], we propose a way to correct significant amount of document preferential errors automatically even if high levels of noise present in the training data.

3. Experimental Setup



4. Noise Injection and Measurement

- Different noise levels are injected on original training data to check efficiency and robustness of noise correction process.
- **Noise Injection:** For noise level p, each pairwise document preference is reversed with probability p and kept the same with probability $1-p$.
- **Noise Measurement:** Fraction of incorrect document preference pairs from the total number of preference pairs in the Post Corrected Document Preferences Set. Correctness is checked compared with Original Document Preferences Set. (Also called *PNoise* [2].)

References

[1] Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A. P., and Yilmaz, E. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (2008), ACM, pp. 667–674.

[2] Niu, S., Lan, Y., Guo, J., Wan, S., and Cheng, X. Which noise affects algorithm robustness for learning to rank. *Information Retrieval Journal 18*, 3 (2015), 215–245.

5. Notation

Partial Pairwise Document Preference Set:

$$\{[\bar{F}(q:d_i>d_j),1] : rel(q,d_i) > rel(q,d_j) \text{ and } d_i,d_j \in D\}$$

Full Pairwise Document Preference Set:

$$\{[\bar{F}(q:d_i>d_j),1] \cup [\bar{F}(q:d_j>d_i),0] : rel(q,d_i) > rel(q,d_j) \text{and} \quad d_i,d_j \in D\}$$

Query	Document	Feature	Relevance
q	d_1	$\bar{f}_{qd_1}$	1
q	d_2	$\bar{f}_{qd_2}$	1
q	d_3	$\bar{f}_{qd_3}$	0
q	d_4	$\bar{f}_{qd_4}$	0

Table 1: **Original Training Data**

Query	Preference	Feature	Correctness
q	$d_1 > d_3$	$\bar{f}_{qd_1} - \bar{f}_{qd_3}$	1
q	$d_1 > d_4$	$\bar{f}_{qd_1} - \bar{f}_{qd_4}$	1
q	$d_2 > d_3$	$\bar{f}_{qd_2} - \bar{f}_{qd_3}$	1
q	$d_2 > d_4$	$\bar{f}_{qd_2} - \bar{f}_{qd_4}$	1

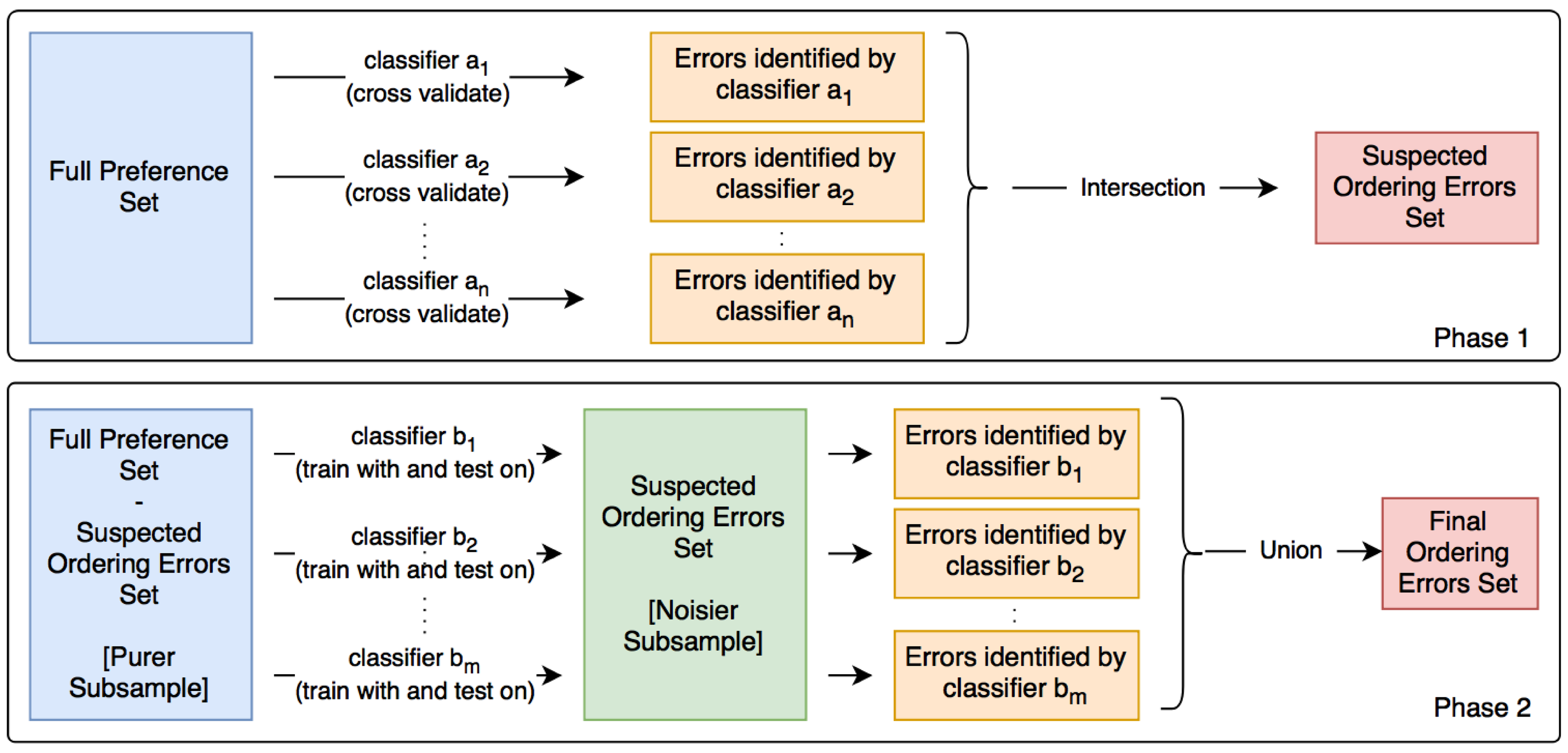
Table 2: **Partial Pairwise Document Preference Set**

Query	Preference	Feature	Correctness
q	$d_1 > d_3$	$\bar{f}_{qd_1} - \bar{f}_{qd_3}$	1
q	$d_1 > d_4$	$\bar{f}_{qd_1} - \bar{f}_{qd_4}$	1
q	$d_2 > d_3$	$\bar{f}_{qd_2} - \bar{f}_{qd_3}$	1
q	$d_2 > d_4$	$\bar{f}_{qd_2} - \bar{f}_{qd_4}$	1

Query	Preference	Feature	Correctness
q	$d_1 > d_3$	$\bar{f}_{qd_1} - \bar{f}_{qd_3}$	0
q	$d_1 > d_4$	$\bar{f}_{qd_1} - \bar{f}_{qd_4}$	0
q	$d_2 > d_3$	$\bar{f}_{qd_2} - \bar{f}_{qd_3}$	0
q	$d_2 > d_4$	$\bar{f}_{qd_2} - \bar{f}_{qd_4}$	0

Table 3: **Full Pairwise Document Preference Set**

6. Two Phase Noise Correction Process



Random Forest, *Multilayerd Perceptron* (weka) classifiers were found to be good choices.

7. Results

Table 4: Noise Correction on TREC-TD-2003

Injected Noise	Post Correction Noise	Percentage Noise Reduction	Queries Improved	Queries Worsened
0.05	0.002	**96.00%***	50	0
0.1	0.006	**94.00%***	50	0
0.15	0.019	**87.33%***	50	0
0.2	0.013	**93.49%***	50	0
0.25	0.023	**90.80%***	50	0
0.3	0.030	**90.00%***	50	0
0.35	0.064	**81.71%***	50	0
0.4	0.108	**73.00%***	50	0
0.45	0.393	**12.66%***	39	11
0.5	0.483	**3.40%**	23	27

Table 5: Noise Correction on TREC-TD-2004

Injected Noise	Post Correction Noise	Percentage Noise Reduction	Queries Improved	Queries Worsened
0.05	0.002	**96.00%***	75	0
0.1	0.004	**96.00%***	75	0
0.15	0.009	**94.00%***	75	0
0.2	0.011	**94.50%***	75	0
0.25	0.027	**89.20%***	75	0
0.3	0.032	**89.33%***	75	0
0.35	0.093	**73.42%***	75	0
0.4	0.109	**72.75%***	75	0
0.45	0.392	**12.88%***	64	11
0.5	0.510	-2.00%	37	38

(*) marked noise reductions are statistically significant.

8. Limitations

- Noise Injection does not model human behaviour.
- Efficacy of successful noise reduction on improvement of performance of Learning to Rank algorithms is not shown.

Acknowledgement

I acknowledge the support of **ACM SIGIR** in the form of a Travel Grant, which enabled me to present this poster at AIRS 2016 conference.