What Makes Data Robust: A Data Analysis in Learning to Rank

Gold Coast SIGIR 2014

Shuzi Niu, Yanyan Lan, Jiafeng Guo, Xueqi Cheng, Xiubo Geng
{niushuzi,gengxiubo}@software.ict.ac.cn, {lanyanyan,guojiafeng,cxq}@ict.ac.cn

Institute of Computing Technology, Chinese Academy of Sciences

Label Noise Problem

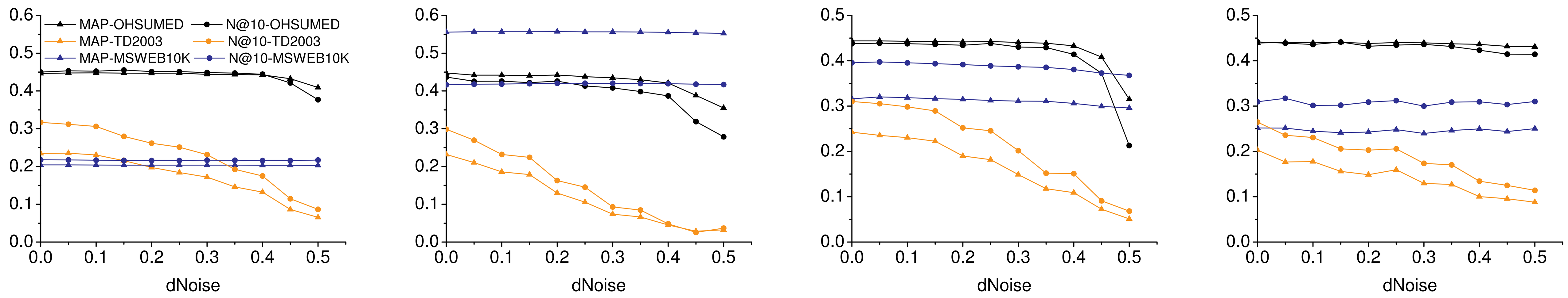
- Noise in human labeled training data will affect the performance of the learning to rank algorithms.
- Related work:
  - How noise affects ranking algorithms;
  - How to design robust ranking algorithms;
  - How to construct effective training data for learning to rank algorithms;
- We focus on what inherent characteristics make training data robust to label noise.

Document Label Noise

- **Document Label Noise** (dNoise): the proportion of documents with error labels to all the documents
- Experiment Setting
  - ∗ Data Sets: OHSUMED, TD2003 and MSLR-WEB10K
  - ∗ Ranking Algorithms: RankSVM, RankBoost, ListNet and AdaRank
  - ∗ Noise Injection Method: randomly change the relevance labels of documents with probability dNoise to other labels uniformly.
  - ∗ For each dNoise, the averaged test performance is obtained by applying this noise injection method 10 times to each training set.

Basic Observation

- A learning to rank algorithm shows very different sensitivities to document label noise over different data sets.



(a) RankSVM (b) RankBoost (c) ListNet (d) AdaRank

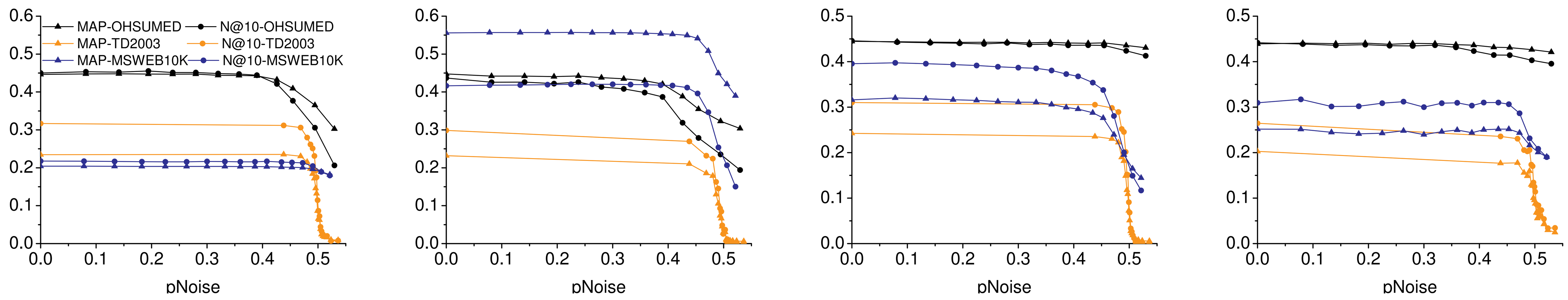
- The above observations are contrary to two intuitions:
  1. *Degradation Intuition*: For a ranking algorithm, its performance would degrade along with the deterioration of the training data quality.
  2. *Consistency Intuition*: For a ranking algorithm, its performance degradation behavior against label noise would be similar across data sets.
- **What is the true noise that affects the performances of ranking algorithms?**

Document Pair Noise

- All document pairs in noisy training set denoted as $V_{all} = \{\langle d_i, d_j\rangle | d_i$ is more relevant than d_j in the noisy training set.$\}$ can be classified into three kinds according to pairwise preferences inferred from relevance labels in the noise free training set:
  - Correct-Order Pairs: $V_{correct} = \{\langle d_i, d_j\rangle | \langle d_i, d_j\rangle \in V_{all}, d_i$ is more relevant than d_j in noise free set.$\}\Rightarrow$**correct**
  - Inverse-Order Pairs: $V_{inverse} = \{\langle d_i, d_j\rangle | \langle d_i, d_j\rangle \in V_{all}, d_i$ is less relevant than d_j in noise free set.$\}\Rightarrow$**harmful**
  - New-Come Pairs: $V_{new} = \{\langle d_i, d_j\rangle | \langle d_i, d_j\rangle \in V_{all}, d_i$ and d_j are of the same label in noise free set.$\}\Rightarrow$**harmful with half probability**
- Document Pair Noise (pNoise)

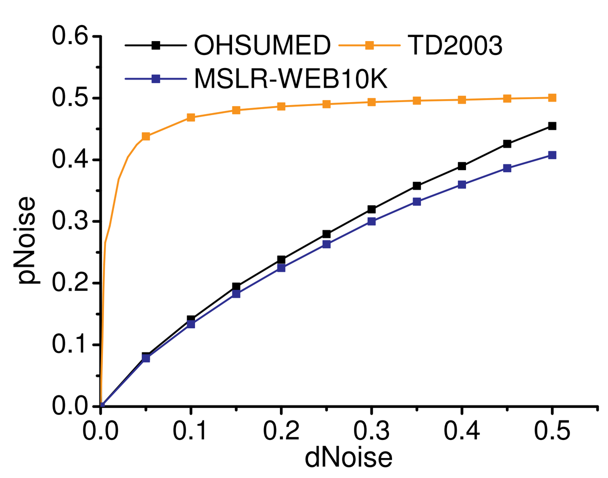
$$\text{pNoise} = \frac{N_{inverse} + 0.5 \times N_{new}}{N_{all}}$$

- The ranking performance variations with respect to pNoise are consistent across different data sets for a ranking algorithm in accordance with the degradation and consistency intuitions.



(a) RankSVM (b) RankBoost (c) ListNet (d) AdaRank

pNoise VS. dNoise



- The increase of pNoise with respect to dNoise varies largely over different data sets.
- This explains the ranking performance variation is different over data sets.
- pNoise reaches its turning point (0.5 or so) on TD2003 more quickly than the other two data sets.

Conclusion

- Conclusion
  - Document pair noise which captures the true noise can well explain the performance degradation of ranking algorithms.
- Future Work
  - Noise injection method can be improved for better analysis.
  - What inherent characteristics of training data have great impact on the variation of pNoise should be investigated further.