Sampling Bias Due to Near-Duplicates in Learning to Rank

Bachelor's Thesis & SIGIR '20 Paper

Jan Heinrich Reimer
jan.reimer@student.uni-halle.de

Supervisor: Maik Fröbe

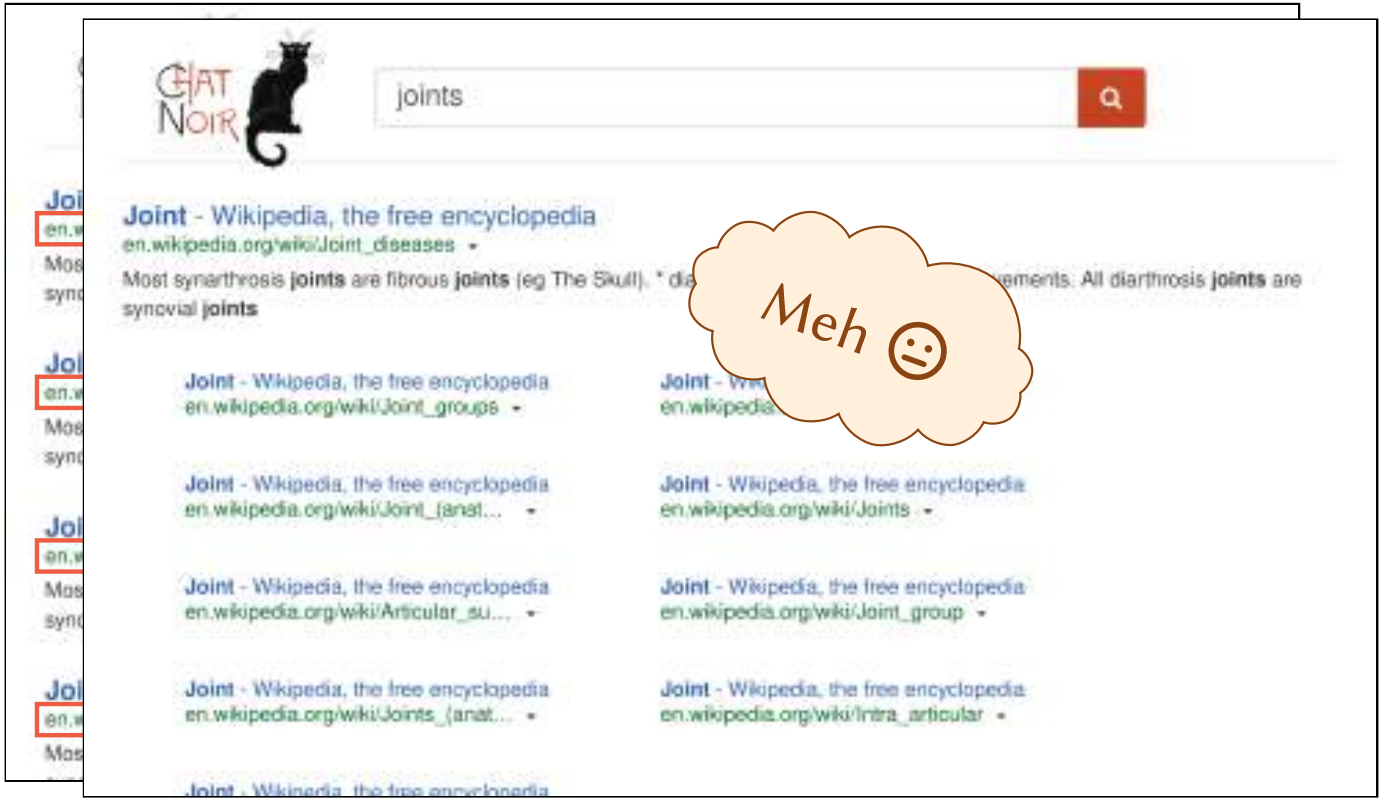
Martin Luther University Halle-Wittenberg

March 12, 2021

Have you been there?

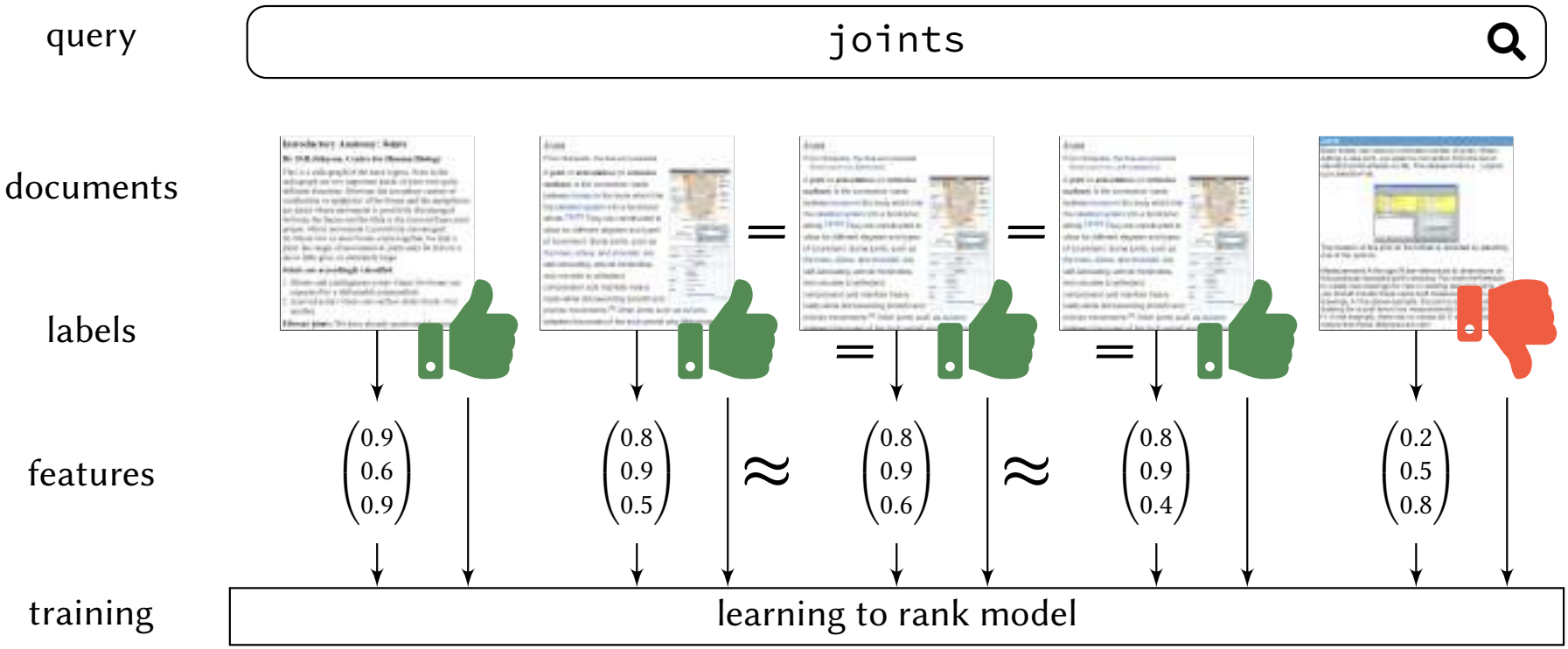
- redundant search results at top ranks

Have you been there?



- redundant search results at top ranks

What's the trouble with Learning to Rank?

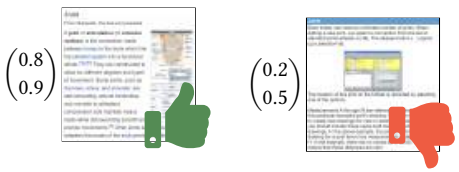


1. identical relevance labels (Cranfield paradigm)
2. similar features, e.g., same TF/IDF
3. oversampling → double impact on loss → overfitting

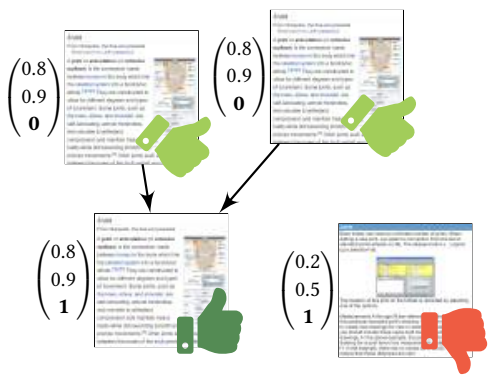
Can we do anything about it?

- reuse methods for counteracting overfitting → undersampling
- canonical link relations [OK12]

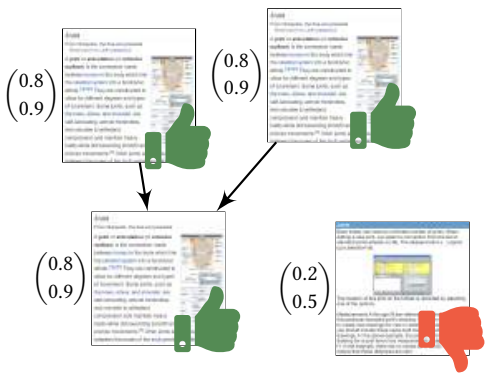
Remove



Discount & flag



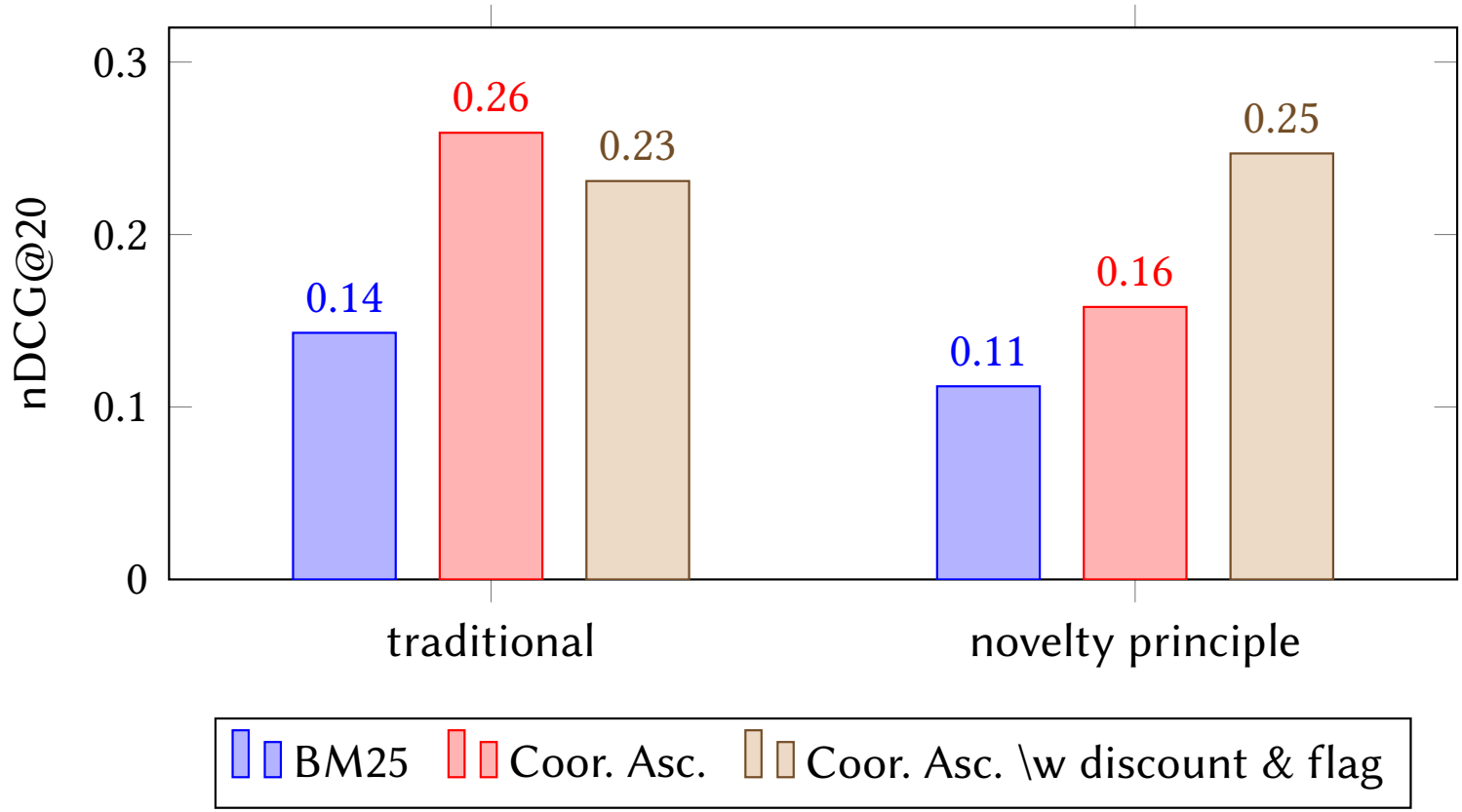
No deduplication



- removing discards training data
- discounting breaks label consistency
- …but works best

How bad is it? Does deduplication work?

Performance for Coordinate Ascent [MC07] on ClueWeb09



- performance decreases under novelty principle [Frö+20]
- discount & flag compensates impact

Conclusion

- near-duplicates reduce retrieval performance in LTR
- **De-duplicate your learning-to-rank training data!**

SIGIR '20 paper

DOI: 10.1145/3397271.3401212

Thank you!