INTRODUCTION

- Electronic discovery (eDiscovery) refers to discovery in civil litigation that deals with the exchange of information in electronic format.
- The corpus may have documents, PDFS, e-mails, audio or video files, and the size of the data that needs to be searched is large.

META-DATA BASED INDEXING

- Indexing facilitates fast and accurate information retrieval.
- Metadata-based indexing facilitates exploring a corpus based on multiple facets.
- We have accomplished this using the Apache Lucene Framework - an open-source information retrieval software library.

METHODS FOR LEARNING-TO-RANK

1. Point-wise:
   - Each query-document pair in the training data has an ordinal score.
   - Learning-to-rank \rightarrow regression problem ((query, document) \rightarrow score prediction)
2. Pair-wise:
   - Learning a binary classifier that can tell which document is better in a given pair of documents.
   - Learning-to-rank \rightarrow classification problem
   - Attempts to minimize the avg. no. of inversions in ranking.
3. List-wise:
   - These algorithms try to directly optimize the value of one of the above evaluation measures, averaged over all queries in the training data.

The algorithms we used in our experimental analysis are:

1. AdaRank:
   - Repeatedly constructs weak learners (learners) on the basis of re-weighted training data.
   - Linearly combines the weak learners for making ranking predictions.
2. ListNet:
   - Has a probabilistic method to calculate the listwise loss function.
   - Transforms both the scores of the documents assigned by a ranking function and the judgments of the documents given by humans into probability distributions.
   - Uses learning to rank method with a listwise loss function, a Neural Network as model, and Gradient Descent as the algorithm.

RESULTS

Comparing learning-to-rank algorithms:

- Shows the comparison between listwise and pairwise ranking algorithms in terms of their performance for Normalized Discounted Cumulative Gain (NDCG).
- Higher the value of NDCG, better is the performance of the ranking algorithm.
- Compares three algorithms -- AdaRank, a listwise algorithm, RankBoost and RankNet, two pairwise algorithms.

The X-axis shows the number of previous documents taken for comparison to calculate NDCG values.

Inferences from our experiments:

- Listwise ranking algorithms have better NDCG values and hence their performance in ranking is better than that of pairwise ranking algorithms.
- Pairwise and listwise algorithms perform better when a higher number of previous documents are taken into consideration while calculating NDCG values.

REFERENCES

1. Zhe, Xu, & Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. (ICML ‘07)

ACKNOWLEDGEMENTS

We thank Dr. Daisy Wang for giving us the opportunity to work on this project. We also thank Clint P. George for his guidance throughout the project.