Feature Extraction for Large-Scale Text Collections

Luke Gallagher, Antonio Mallia, J. Shane Culpepper, Torsten Suel, and B. Barla Cambazoglu

RMIT University

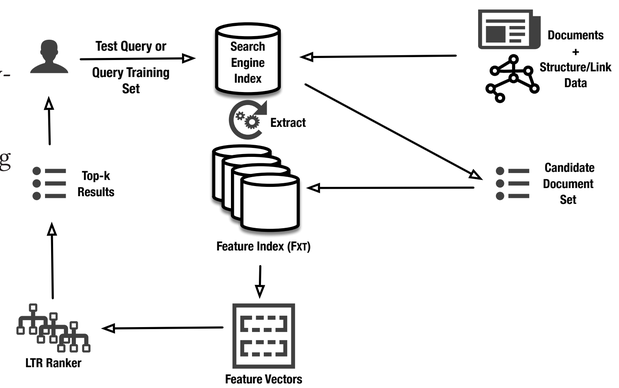
NYU

1. Contributions: Feature Extraction Toolkit and Dataset

- Feature engineering is a fundamental component for many Learning-to-Rank (LTR) search applications Although we rarely see open and accessible documentation and/or software for researching and implementing feature extraction capabilities in the context of a large-scale search system
- This work provides two resources that aim to make it easier when working on problems related to feature extraction and feature-based search models
  1. **Feature Extraction Toolkit**—Software that helps facilitate feature extraction processes within search tasks.
  2. **LTR Dataset**—An open and transparent dataset built on ClueWeb09B with the feature extraction toolkit.

2. Feature Extraction Toolkit

- Configurable collection of 448 features
- Features mainly derived from the literature on Query-Performance Prediction and LTR.
- The software has two main components for indexing and the extraction of features.
- Some example tasks are:
  - Build an index optimized for feature extraction
  - Standalone feature extraction
  - Generate training data



3. Learning-to-Rank Dataset

- Dataset derived from ClueWeb09B using the Million Query and Web Tracks from 2009–2012.
- BM25 was used to generate candidate documents, and query sets were ranomdly shuffled.
- Includes 134 features
- Different relevance grades and judging methods were used over the years leading to the train-test query sets shown.

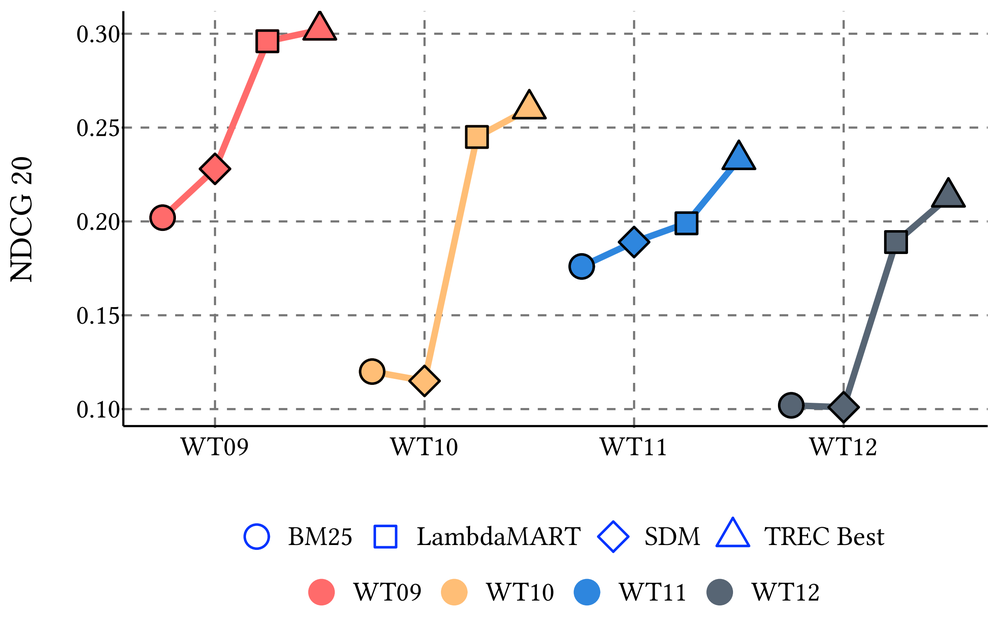
Test Queries	Train/Valid Queries
WT09	MQ09
WT10	WT09, WT11, WT12
WT11	WT09, WT10, WT12
WT12	WT09, WT10, WT11

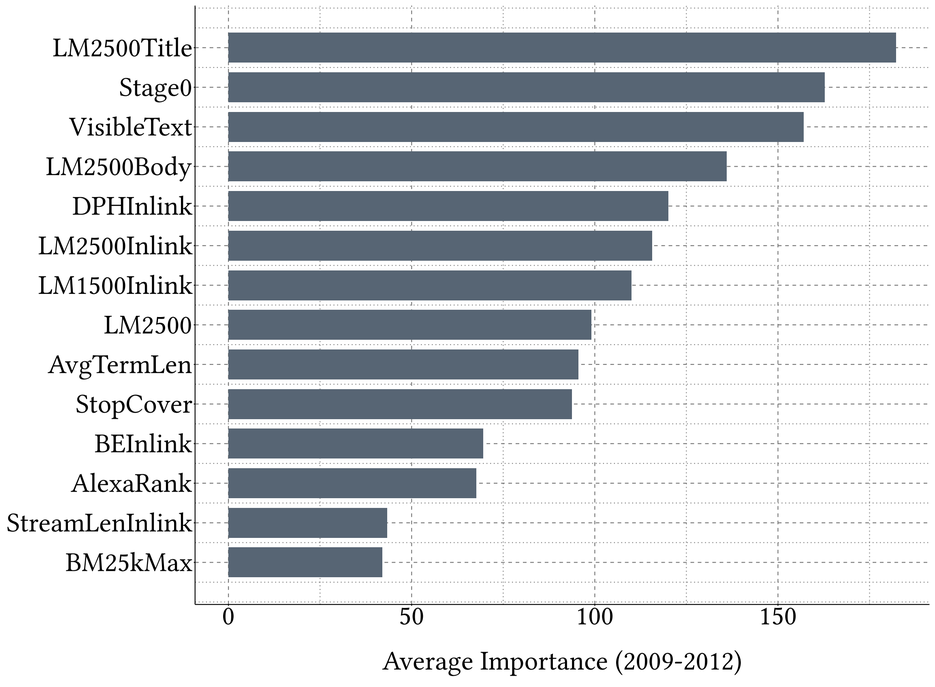
4. Experimental Setup

- Compare effectiveness of LambdaMART using LightGBM, against traditional baselines
- Evaluation on the 4 Web Track query sets
- Conduct brief study on feature importance
- A summary of the features used in the dataset are shown

Feature Class	No. Features
Query-Document (Unigram)	106
Query-Document (Bigram)	4
Static Document	23
AlexaRank (Not available in Fxt)	1

5. Results and Analysis





Acknowledgements. This work was supported by the Australian Research Council's *Discovery Projects* Scheme (DP190101113), the NSF Grant IIS-1718680, an Amazon Research Award, and an Australian Government Research Training Program Scholarship.