Selective Search: a distributed search architecture using topical shards to reduce computational costs.

1. Sample
- Topical Shards
- BUILD RESOURCE SELECTION
- Project remaining documents
- BUILD

Resource Selection:
- Rank-S
- Taily
- RBR: ranking-based resource selection.
- An oracle algorithm.

Two Sources of Errors
1. Partitioning
- The only error of RBR comes from partitioning.
- RBR box-plots show that some partitioning results were better than the others.

2. Resource Selection
- Box-plots of Rank-S were wider than RBR, and had more outliers.
- Resource selection algorithms introduced additional errors due to incomplete models.

Variability of Query Sets

Experiment Setup
- ClueWeb 09 (Category A, B)
- 4 query sets: TREC 2009-2012
- Resource Selection:
  - Rank-S, Taily
  - RBR: ranking-based resource selection.
- An oracle algorithm.
- 10 Trials

Two Sources of Errors

1. Partitioning
- The only error of RBR comes from partitioning.
- RBR box-plots show that some partitioning results were better than the others.

2. Resource Selection
- Box-plots of Rank-S were wider than RBR, and had more outliers.
- Resource selection algorithms introduced additional errors due to incomplete models.

Sample-Based vs. Vocabulary-Based
- Rank-S: 3 random decisions
- 2 in partitioning: sampling + clustering
- 1 in resource selection: sampling
- Taily: 2 random decisions
- 2 in partitioning: same as Rank-S
- NO randomness in resource selection

Which is more stable, Rank-S or Taily?

Variability of Queries

Most queries are stable.
- # of high variance queries:
  - Rank-S: 14 out of 200
  - Taily: 17 out of 200

Variance mainly comes from partitioning.
- High-variance queries in Rank-S and Taily also had high variance in RBR.
- RBR does not have resource-selection errors. Therefore, the major source of variance comes from partitioning.

What is ‘poor’ partitioning?
- ❌ Is ‘poor’ partitioning due to relevant documents being grouped into more shards?
- NO. All trials grouped over 60% of relevant documents into 3 or fewer shards.
- ✅ A ‘poor’ partitioning groups relevant documents with topically dissimilar documents.
- Shards are unrepresentative of relevant documents, making resource selection harder.