How Random Decisions Affect Selective Distributed Search

Zhuyun Dai, Yubin Kim, Jamie Callan
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{zhuyund, yubink, callan}@cs.cmu.edu

ABSTRACT

Selective distributed search is a retrieval architecture that reduces search costs by partitioning a corpus into topical shards such that only a few shards need to be searched for each query. Prior research created topical shards by using random seed documents to cluster a random sample of the full corpus. The resource selection algorithm might use a different random sample of the corpus. These random components make selective search non-deterministic.

This paper studies how these random components affect experimental results. Experiments on two ClueWeb09 corpora and four query sets show that in spite of random components, selective search is stable for most queries.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval

Keywords

distributed retrieval, selective search, variance

1. INTRODUCTION

A *selective search* architecture reduces search costs by organizing a large corpus into topical index *shards* and searching only the most likely shards for each query. As defined by prior research, selective search has several non-deterministic steps. Kulkarni [4] used sample-based k-means clustering to create topical shards, which is non-deterministic due to random sampling and cluster seeds. Some resource selection algorithms, such as Rank-S [5], use a random sample of the corpus to decide which shards to search for each query, which introduces additional non-determinism.

Prior research showed that selective search can be as accurate as a typical 'search all shards' distributed architecture but at a substantially reduced computational cost [1, 4, 5]. However, it was based on a single partitioning of the corpus. A different partitioning might yield different results.

We are not the first to notice this problem. Jayasinghe, et al. [3] proposed a linear model to statistically compare non-deterministic retrieval systems, using selective search as an example. Their model took into consideration the multi-dimensional variance in selective search. However, their work focused on testing whether selective search has equivalent mean effectiveness as a baseline search architecture. In our work, we focus on the variance itself.

This paper investigates the effect of random decisions on selective search accuracy by comparing results obtained with different partitionings of two ClueWeb datasets. It examines the variance of a system across four query sets (Section 3.1), and for individual queries (Section 3.2).

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

ACM 978-1-4503-3621-5/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767796 .

2. EXPERIMENTAL METHOD

We define a *system instance* to be a partition of a corpus index and its corresponding resource selection database for algorithms such as Rank-S [5] or Taily [1]. Our system instances were defined by a slight adaptation of a process defined by Kulkarni [4]. First, 500K documents were randomly sampled from the corpus, k of those documents were selected to be seeds, and k-means clustering was used to form clusters. Second, the remaining documents were projected onto the most similar clusters to form index shards. Third, spam documents were removed and a resource selection index was constructed. The Rank-S resource selection index was formed from a 1% random sample of each cluster. The Taily resource selection index is created deterministically. Thus, each system instance involved 2-3 random processes.

The parameters used for Taily and Rank-S were suggested by Aly et al. [1] and Kulkarni et al. [5]. They were $n = 400$, $v = 50$ for Taily and $B = 5$, $CSI = 1\%$ for Rank-S.

The experiments also tested Relevance-Based Ranking (RBR), an oracle resource selection algorithm that ranks shards by the number of relevant documents that they contain. RBR makes it possible to distinguish between different types of variance. For query q, RBR searched the average number of shards that Rank-S and Taily searched for q.

Experiments were conducted on the ClueWeb09 Category A (*CW09-A*) corpus, which contains 500 million English web pages, and ClueWeb09 Category B (*CW09-B*), which contains 50 million English web pages. Queries were from the TREC 2009-2012 Web Tracks: 4 sets of 50 queries.

Parameters were set to produce shards of 500K documents on average. Each CW09-A system instance had 1000 shards; CW09-B system instances had 100 shards.

3. EXPERIMENT RESULTS

The effects of random decisions during partitioning and indexing can be measured across query sets or individual queries. Measuring across query sets provides information

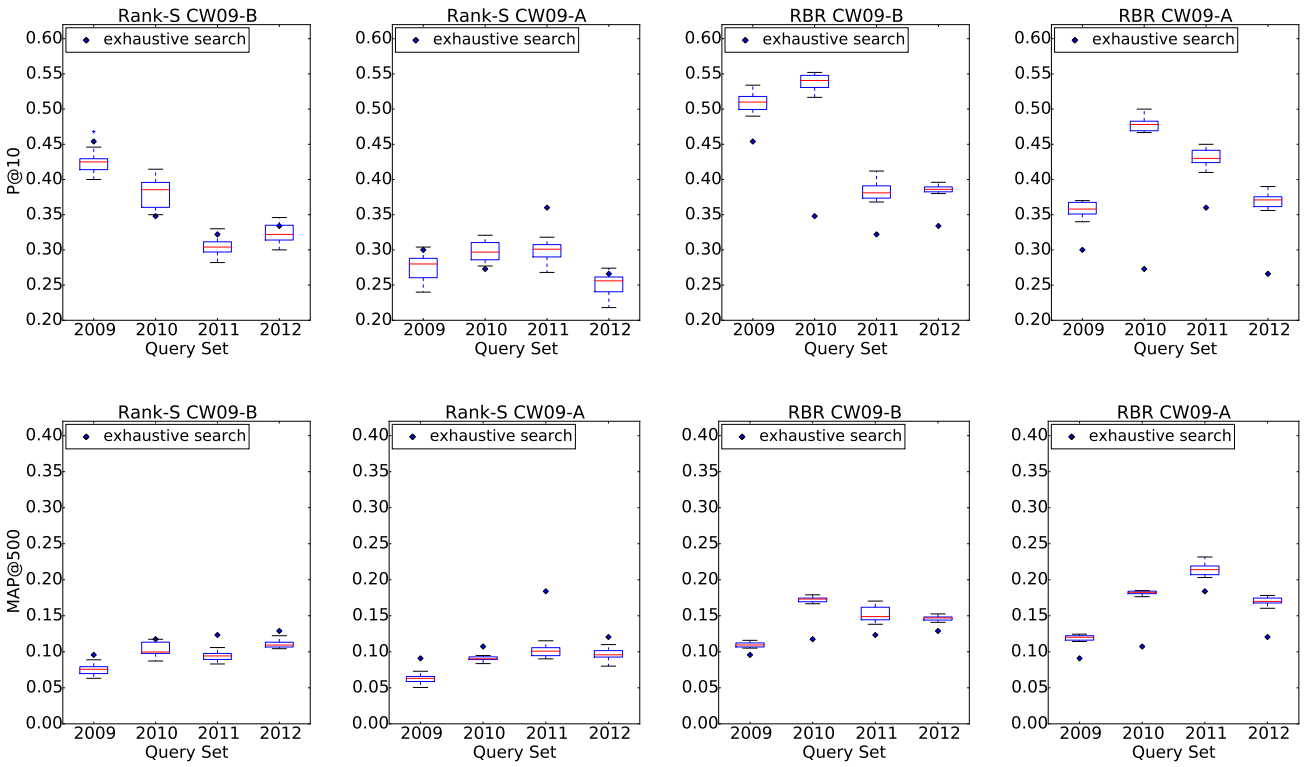

Figure 1: Accuracy distribution of Rank-S and RBR system instances. Mid line in boxes represents median. Outer edges of the boxes represent lower and upper quartiles, while whiskers indicate 1.5 interquartile range. + indicates outliers. Blue diamond is the performance of the typical 'search all shards baseline.

about the reliability of metrics such as MAP and NDCG that are typically reported in research papers. Measuring across individual queries provides information about the behavior that people will observe when they use the system. The experiments below provide both types of measurement.

3.1 Variability for Query Sets

The first experiment examined how random decisions affect the 'average case' metrics typically reported in research papers. It used three resource selection algorithms: Rank-S, which has a random component; Taily, which is deterministic; and RBR, the oracle algorithm. Ten random partitions were created for each dataset. Thus there were 2 datasets $\times$ 10 random partitions $\times$ 3 resource selection algorithms $=$ 60 system instances. On average, Rank-S searched 3.91 CW09-A shards and 3.75 CW09-B shards; Taily searched 2.58 CW09-A shards and 2.53 CW09-B shards; and RBR searched 3.13 CW09-A shards and 2.80 CW09-B shards.

Retrieval accuracy was evaluated using three metrics: Precision at rank 10 (P@10), Normalized Discounted Cumulative Gain at rank 100 (NDCG@100), and Mean Average Precision at rank 500 (MAP@500) for compatibility with prior research [1]. Results for Taily and NDCG@100 were similar to results for Rank-S and other metrics, thus they are omitted due to space limitations.

The distributions of measurements over each type of system instance and the 'search all shards' baseline are shown in Figure 1. Tables 1 and 2 report the standard deviation and variance coefficient of Rank-S and RBR system accuracy scores. The standard deviation is a common measure, however it has two flaws: i) values produced by different metrics (e.g., P@10, MAP@500) cannot be compared because it is not dimensionless; and ii) the *importance* of a given level of deviation depends upon the value of the mean. The *variance coefficient* (VC) normalizes the standard deviation by the mean, thus eliminating these flaws.

The Relevance-Based Ranking (RBR) was generally more accurate and had lower variance across system instances than Rank-S and Taily, as expected. The one exception was the 2011 query set and the MAP@500 metric; Rank-S had lower variance than RBR under this condition. This exception was caused by a small set of queries. Rank-S had consistently poor accuracy on these queries, whereas RBR had greater accuracy but also greater variance.

The experimental results suggest that there are two sources of variance in selective search: Partitioning, and resource selection. RBR always selects the best shards, thus variance in RBR results is due to differences in partitioning effectiveness. Rank-S and Taily use models (resource selection indices) to make decisions about which shards to select for each query, however those models are necessarily incomplete. The increases in variance for Rank-S and Taily as compared to RBR are due to resource selection errors caused by weaknesses in the models of shard contents.

Selective search instances displayed differing levels of variability across different metrics. P@10 overall had a lower variance coefficient than MAP@500 (Tables 1 and 2). Similar behavior is observed for Taily instances. This behavior

Table 1: Standard Deviation (SD) and Variance Coefficient (VC) of accuracy scores for Rank-S instances.

a) P@10

Query Set	CW09-B		CW09-A	
	SD	VC	SD	VC
2009	1.85×10^{-2}	4.34%	2.00×10^{-2}	7.30%
2010	2.14×10^{-2}	5.61%	1.50×10^{-2}	4.90%
2011	1.43×10^{-2}	4.71%	1.51×10^{-2}	5.06%
2012	1.49×10^{-2}	4.60%	1.54×10^{-2}	6.14%

b) MAP@500

Query Set	CW09-B		CW09-A	
	SD	VC	SD	VC
2009	8.26×10^{-3}	10.79%	6.65×10^{-3}	10.65%
2010	9.66×10^{-3}	9.34%	3.31×10^{-3}	3.66%
2011	6.68×10^{-3}	7.14%	7.67×10^{-3}	7.62%
2012	5.42×10^{-3}	4.90%	7.96×10^{-3}	8.23%

indicates that selective search is more stable at the top of the ranking.

Rank-S is affected by one more random component than Taily, thus it might be expected to have greater variability across system instances. Table 3 shows the variance coefficient of MAP@500 for Rank-S and Taily across ten CW09-A system instances; similar behavior was observed across all metrics for both collections. Taily only had lower variance than Rank-S on 1 of the 4 query sets. These results might be considered surprising. The additional random component does not appear to cause greater instability in Rank-S results. As far as we know, prior research has not investigated the variability of results produced by document-based and model-based algorithms such as Rank-S and Taily.

3.2 Variability for Queries

Ideally selective search would provide similar results for a given query regardless of which system instance is used. The second experiment examined variability on a query-by-query basis.

We examined the variance of Average Precision of each query across Rank-S and Taily instances on CW09-A and observed some highly variant queries in both Rank-S and Taily. Rank-S had 14 queries with AP standard deviation higher than 0.1; Taily had 17 such high variance queries.

High variance could be due to errors from the partitioning process or errors from resource selection. To discover the source of the variance, a query-by-query experiment was conducted with RBR on the same partitions. RBR does not make resource selection errors but *is* affected by partitioning errors. Thus, if Rank-S, Taily, and RBR all have trouble with a query, the problem is likely to be partitioning.

Figure 2 shows the average precision (AP) of each query for ten RBR system instances. A notable observation is that most of the queries have stable AP; 177 of the 200 queries had a standard deviation of less than 0.05. However, there were a few queries with very high variance that contributed most of the average variance.

Whether the query is variable or stable remained mostly consistent across RBR, Rank-S, and Taily. While there were some highly variable queries in RBR that were not so in Taily and Rank-S, these were difficult queries (low 'search all shards' search accuracy), and had nearly zero MAP@500

Table 2: Standard Deviation (SD) and Variance Coefficient (VC) of accuracy scores for RBR instances.

a) P@10

Query Set	CW09-B		CW09-A	
	SD	VC	SD	VC
2009	1.38×10^{-2}	2.71%	1.05×10^{-2}	2.95%
2010	1.16×10^{-2}	2.15%	9.73×10^{-3}	2.04%
2011	1.36×10^{-2}	3.53%	1.31×10^{-2}	3.04%
2012	5.39×10^{-3}	1.39%	9.73×10^{-3}	2.63%

b) MAP@500

Query Set	CW09-B		CW09-A	
	SD	VC	SD	VC
2009	3.56×10^{-3}	3.25%	3.54×10^{-3}	2.97%
2010	3.87×10^{-3}	2.24%	2.46×10^{-3}	1.35%
2011	1.10×10^{-2}	7.21%	9.47×10^{-3}	4.40%
2012	3.53×10^{-3}	2.41%	4.97×10^{-3}	2.91%

Table 3: Comparison of MAP@500 scores (CW09-A) for Rank-S and Taily instances.

Query Set	Mean		Variance Coefficient	
	Rank-S	Taily	Rank-S	Taily
2009	0.062	0.065	10.65%	10.98%
2010	0.090	0.087	3.66%	6.89%
2011	0.101	0.084	7.62%	15.67%
2012	0.097	0.085	8.23%	5.36%

in Rank-S and Taily. All easy queries with high variance in RBR also had high variance in Rank-S and Taily. The consistency across RBR, Rank-S and Taily suggests that errors from the partitioning stage are a major source of variance of selective search accuracy; however, most of that variation comes from a small number of queries.

We investigated why some partitions performed much better than others on the high variance queries. One might expect that 'good' partitions group relevant documents into fewer shards. However, in all of our partitions, relevant documents were distributed across a small number of shards. Typically the 3 most relevant shards contained more than 60% of the relevant documents, which is consistent with the standards of most prior federated search research. Furthermore, in this experiment, all ten instances retrieved 100% of the relevant documents for 4 of the 5 most variant queries. However, these queries had AP values ranging from values similar to exhaustive search to as much as ten times *better* than exhaustive search.

An examination of the instances with unusually high AP in the RBR experiment revealed that the relevant documents and the most likely false positives were in different shards. Thus, the combination of partitioning and resource selection 'filtered out' non-relevant documents that otherwise would have been ranked highly. Although this behavior might seem desirable, it occurred because the partitioning process incorrectly assigned relevant documents to shards that contained documents that were largely dissimilar to the relevant documents. This poor clustering made it nearly impossible for Rank-S and Taily to identify which shards contained relevant documents due to the overwhelming number of dissimilar documents in the shard. RBR, which *knows* the number of relevant documents in each shard, had no such problem.

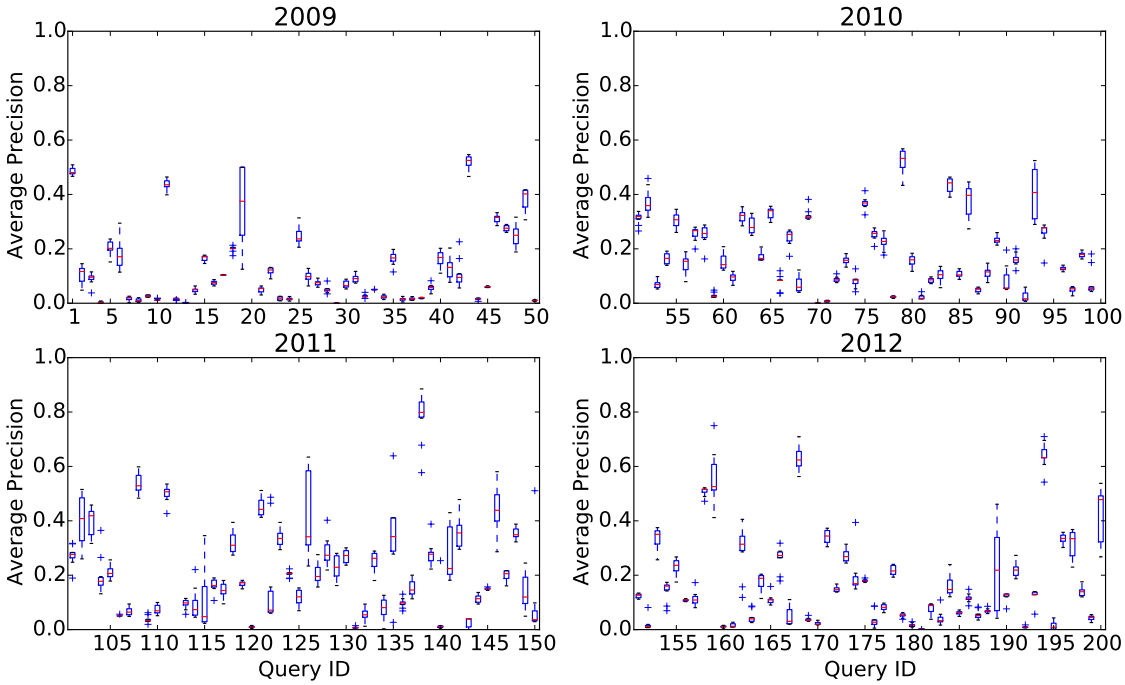

Figure 2: Query Average Precision of RBR instances. Each data point is a box and whisker plot with the edges of the boxes representing the upper and lower quantiles, the mid line the median, and the whiskers the 1.5 interquartile range. + are outliers.

Figure 2 also indicates that different query sets produced different levels of variability. The 2011 query set was notably more unstable than the others. Among the 24 queries with AP standard deviation higher than 0.05, 14 were from TREC 2011. We believe that this difference is due to the other three query sets using topics of medium-to-high frequency, while TREC 2011 used more obscure topics [2]. Our results indicate that the *type* of query has an impact on not only the average accuracy of a system, but also the variance of a system with random components. Topic-based partitioning may produce stable results for common topics, but may need improvement for queries about rare topics.

4. CONCLUSION

Understanding the variance of selective search effectiveness is critical for evaluating and improving selective search. This paper explores selective search variance and the effects of document collections, resource selection algorithms, and query characteristics. Partitioning and resource selection processes both introduce variance into search behavior, however in our experiments the effect on most queries was not large. Results were more stable at the top of the ranking, but variance at rank 500 was not unreasonable, especially for an architecture that avoids searching most of the index.

Rank-S and Taily produced nearly equal variance, which might be considered surprising given that Rank-S has one more random component than Taily.

Most of the variance observed in our per-query experiments was caused by a small number of queries – typically, rare queries. In some instances the partitioning process correctly grouped relevant documents together, but placed them in unrepresentative shards, which caused poor resource selection. This behavior may indicate the need for partitioning processes more sophisticated than simple k-means clustering.

5. ACKNOWLEDGMENTS

This research was supported by National Science Foundation (NSF) grant IIS-1302206 and Natural Sciences and Engineering Research Council of Canada (NSERC) Postgraduate Scholarship-Doctoral. Any opinions, findings, conclusions, and recommendations expressed in this paper are the authors' and do not necessarily reflect those of the sponsors.

6. REFERENCES

[1] R. Aly, T. Demeester, and D. Hiemstra. Taily: Shard selection using the tail of score distributions. In *Proceedings of SIGIR*, 2013.

[2] C. L. Clarke, N. Craswell, I. Soboroff, and E. M. Voorhees. Overview of the trec 2011 web track. In *Proceedings of TREC 2011*, 2011.

[3] G. K. Jayasinghe, W. Webber, M. Sanderson, L. S. Dharmasena, and J. S. Culpepper. Evaluating non-deterministic retrieval systems. In *Proceedings of SIGIR*, 2014.

[4] A. Kulkarni. *Efficient and Effective Large-scale Search.* PhD thesis, Carnegie Mellon University, 2013.

[5] A. Kulkarni, A. Tigelaar, J. Callan, and D. Hiemstra. Shard ranking and cutoff estimation for topically partitioned collections. In *Proceedings of CIKM*, 2012.