decreases because e in Figure 1 decreases (i.e. cases where both AUReC and end-to-end believe the differences between the two shard maps are not significant).

We have calculated AUReC with queries without relevance judgements or high accuracy runs, using an out-of-the-box search engine. The results were highly correlated with full end-to-end system evaluations. When this less optimal configuration is compared to the optimal cluster effectiveness baseline from Section 5.1 which used the same queries and relevance judgements as the end-to-end system evaluations, AUReC had better pair-wise correlation and better or near-equal list-wise correlation while using a different set of queries and no relevance judgements, demonstrating that it is robust and reliable.

Num of qrys	r	Pairs recall	Overlap+Addit. pairs
100	0.89	186/236 = 0.79	0.79 + 0.08 = 0.86
400	0.92	202/236 = 0.86	0.77 + 0.12 = 0.89
800	0.93	217/236 = 0.92	0.70 + 0.21 = 0.91
1000	0.93	219/236 = 0.93	0.66 + 0.24 = 0.90
10000	0.93	224/236 = 0.95	0.58 + 0.28 = 0.86

(a) ClueWeb09 B

Num of qrys	r	Pairs recall	Overlap+Addit. pairs
100	0.89	178/226 = 0.79	0.72 + 0.12 = 0.84
400	0.93	211/226 = 0.93	0.74 + 0.16 = 0.90
800	0.95	218/226 = 0.96	0.71 + 0.20 = 0.91
1000	0.94	218/226 = 0.96	0.69 + 0.21 = 0.90
10000	0.96	225/226 = 1.00	0.58 + 0.28 = 0.86

(b) Gov2

Table 5: Comparsion of AUReC and Rank-S end-to-end system evaluation using P@1000, when using varying number of MQT queries. The end-to-end system was evaluated with TREC queries, as usual.

6 CONCLUSION AND RECOMMENDATIONS

Prior work evaluated shard maps by measuring the accuracy of end-to-end selective search systems. This is a cumbersome method that relies on relevance judgements and is sensitive to the specific system configuration. This paper introduces AUReC, a new way to measure the effectiveness of shard maps that does not require gathering relevance judgments and is the first to completely decouple shard map evaluation from other components and parameters of a selective search system. By freeing shard map quality from other system components, AUReC provides robust diagnostic information that can be used to quickly sort through a large number of shard maps to tune a new selective search system, a process which was previously time-consuming and difficult.

AUReC evaluates shard maps by the area under a recall curve using the retrieval results of an exhaustive search system. It is highly-correlated to end-to-end selective search system evaluations while being simple to implement and not requiring: the implementation of other selective search components; picking a fixed efficiency level; or human-assessed relevance judgements. An examination of the effectiveness and robustness of AUReC found it produces scores that are highly-correlated with the evaluation of end-to-end systems under a variety of configurations.

Given a set of shard maps, the ordering of the shard maps determined by AUReC scores closely resembled the ordering by different end-to-end evaluations, usually with Pearson's $r > 0.9$. When pairs of shard maps were compared, most shard maps that had significant differences under an end-to-end evaluation also were significantly different when compared with AUReC scores. AUReC scores are calculated from easy-to-generate, plentiful data points and therefore produces stable results. Thus, AUReC was able to ascertain significant differences in pairs of shard maps where end-to-end system evaluations could not due to the scarcity of relevance data.

AUReC allows system designers to quickly test a large number of shard maps to tune the accuracy of a new selective search system, a task which used to be prohibitively expensive. We end the paper with practical guidelines on using AUReC to tune a system. First, to generate D_q, the set of documents that should be retrieved for query q, the strongest search engine available is preferred. However, an out-of-the-box retrieval still produces reliable results. More queries generate more consistent results with less variance. However, there are diminishing gains after about 800 queries.

REFERENCES

[1] Robin Aly, Djoerd Hiemstra, and Thomas Demeester. 2013. Taily: Shard Selection Using the Tail of Score Distributions. In *Proceedings of SIGIR*. 673–682.
[2] Yael Anava, Anna Shtok, Oren Kurland, and Ella Rabinovich. [n. d.]. A Probabilistic Fusion Framework. In *Proceedings of CIKM*.
[3] Ulf Brefeld, B. Barla Cambazoglu, and Flavio P. Junqueira. 2011. Document Assignment in Multi-site Search Engines. In *Proceedings of WSDM*. 575–584.
[4] B. Barla Cambazoglu, Emre Varol, Enver Kayaaslan, Cevdet Aykanat, and Ricardo Baeza-Yates. 2010. Query Forwarding in Geographically Distributed Search Engines. In *Proceedings of SIGIR*. 90–97.
[5] David Carmel, Doron Cohen, Ronald Fagin, Eitan Farchi, Michael Herscovici, Yoelle S. Maarek, and Aya Soffer. 2001. Static Index Pruning for Information Retrieval Systems. In *Proceedings of SIGIR*. 43–50.
[6] Charles L. A. Clarke, J. Shane Culpepper, and Alistair Moffat. 2016. Assessing efficiency–effectiveness tradeoffs in multi-stage retrieval systems without using relevance judgments. *Inf. Ret.* 19, 4 (2016), 351–377.
[7] Zhuyun Dai, Yubin Kim, and Jamie Callan. 2015. How Random Decisions Affect Selective Distributed Search. In *Proceedings of SIGIR*. 771–774.
[8] Zhuyun Dai, Yubin Kim, and Jamie Callan. 2017. Learning To Rank Resources. In *Proceedings of SIGIR*. 837–840.
[9] Zhuyun Dai, Chenyan Xiong, and Jamie Callan. [n. d.]. Query-Biased Partitioning for Selective Search. In *Proceedings of CIKM*.
[10] James C. French and Allison L. Powell. 2000. Metrics for evaluating database selection techniques. *World Wide Web* 3, 3 (2000), 153–163.
[11] Alan Griffiths, H.Claire Luckhurst, and Peter Willett. 1986. Using Inter-document Similarity Information in Document Retrieval Systems. *J. Am. Soc. Inf. Sci.* 37 (1986), 3–11.
[12] Yubin Kim, Jamie Callan, J. Shane Culpepper, and Alistair Moffat. 2017. Efficient distributed selective search. *Inf. Ret.* 20, 3 (2017), 221–252.
[13] Anagha Kulkarni and Jamie Callan. 2010. Document Allocation Policies for Selective Searching of Distributed Indexes. In *Proceedings of CIKM*. 449–458.
[14] Anagha Kulkarni, Almer Tigelaar, Djoerd Hiemstra, and Jamie Callan. 2012. Shard Ranking and Cutoff Estimation for Topically Partitioned Collections. In *Proceedings of CIKM*. 555–564.
[15] Xiaolu Lu, Alistair Moffat, and J. Shane Culpepper. 2016. The Effect of Pooling and Evaluation Depth on IR Metrics. *Inf. Ret.* 19, 4 (2016), 416–445.
[16] Craig Macdonald, Rodrygo L. T. Santos, and Iadh Ounis. 2013. The whens and hows of learning to rank for web search. *Inf. Ret.* 16, 5 (2013), 584–628.
[17] Luo Si and Jamie Callan. 2003. Relevant Document Distribution Estimation Method for Resource Selection. In *Proceedings of SIGIR*. 298–305.
[18] Anastasios Tombros, Robert Villa, and C.J Van Rijsbergen. 2002. The effectiveness of query-specific hierarchic clustering in information retrieval. *Inf. Process. Manage.* 38, 4 (2002), 559–582.
[19] Ellen M. Voorhees. 1985. *The Effectiveness and Efficiency of Agglomerative Hierarchic Clustering in Document Retrieval.* Technical Report. Cornell University.