The Impact of Collection Size on Relevance and Diversity

Marijn Koolen Jaap Kamps

University of Amsterdam, The Netherlands
{m.h.a.koolen, kamps}@uva.nl

ABSTRACT

It has been observed that precision increases with collection size. One explanation could be that the redundancy of information increases, making it easier to find multiple documents conveying the same information. Arguably, a user has no interest in reading the same information over and over, but would prefer a set of diverse search results covering multiple aspects of the search topic. In this paper, we look at the impact of the collection size on the relevance and diversity of retrieval results by down-sampling the collection. Our main finding is that we can we can improve diversity by randomly removing the majority of the results—this will significantly reduce the redundancy and only marginally affect the subtopic coverage.

Categories and Subject Descriptors: H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*performance evaluation (efficiency and effectiveness)*

General Terms: Experimentation, Measurement, Performance

Keywords: Diversity, Relevance, Collection size

1. INTRODUCTION

Hawking and Robertson [2] established that precision at a certain rank cutoff increases as the collection grows in size. Other things being equal, a larger collection will contain more relevant documents making it intuitively easier to find a fixed number of them. Hence we postulate that:

1. The amount of *relevant information* increases with collection size.

However, adding documents to the collection will lead to diminishing returns: since more and more information is already covered by the collection, it is increasingly hard to add new information. Hence we postulate that:

2. The amount of *redundant information* increases with collection size.

The TREC 2009 Web Track's Diversity Task [1] addresses the issue of redundancy by penalising systems that return the same information over and over again. Diversity puts the impact of collection size on precision in an interesting perspective. On the one hand, being topically relevant is a prerequisite for the desired results, which according to our first postulate would make a larger collection size beneficial. On the other hand, redundancy of information is harmful, which according to our second postulate would make a larger collection potentially detrimental. We will try to determine the relative importance of these two opposing forces. Hence our main research question is:

- ▷ What is the impact of collection size on the diversity of search results?

We use ClueWeb09 category B, consisting of the first 50 million English pages of the full ClueWeb09 collection and the Diversity Task's topics and relevance judgements. We indexed the collection using Indri 4.10. Common stop words are removed and the remaining terms are stemmed using Krovetz. The retrieval model is a standard language model with Jelinek-Mercer smoothing ($\lambda = 0.15$) and a linear length prior (proportional to the length of the document). This run is not optimised for diversity, but merely serves as a way to illustrate the phenomena under consideration. We randomly down-sample the collection, using collection samples ranging between 5% and 100% of the full collection, and repeat this experiment five times. All sample statistics and scores are averages over these five iterations. Random sampling will make the expected probability of relevance of a document the same in the sample and in the full collection. This is helpful for our analysis, but in a realistic setting collections are unlikely to grow in an unbiased way.

2. RELEVANCE AND COVERAGE

We will first analyse the effect of reducing the collection size on the number of relevant documents, and on the number of topics or subtopics with at least one relevant result. There are 50 Diversity topics with 180 subtopics having at least one relevant page in the ClueWeb09 category B collection. In total, there are 3,777 positive relevance judgments for 2,783 distinct pages (some pages are relevant for multiple subtopics). Figure 1 shows the fraction of relevant pages in each sample and the fraction of subtopics for which there is at least one relevant page in the sample (averaged over the five samples). What is the impact of collection size on the number of relevant documents? Obviously, with random sampling the fraction of the relevant pages increases proportionally with the collection. Our first postulate holds.

What is the impact on the number of topics or subtopics with at least one relevant result? Here we see a very different pattern. Starting at 5%, the sample already contains over 40% of the subtopics. At a sample size of 30%, the collection contains relevant pages for over 80% of the subtopics. The fractions for the overall topics are even higher.

Our analysis shows that the small samples already cover the vast majority of subtopics with a relatively small fraction of the relevant documents. The larger samples, in contrast, contain many more relevant documents but only few additional subtopics.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

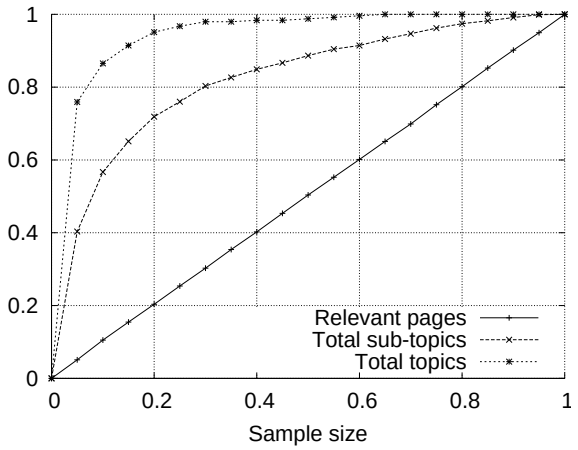

Figure 1: Impact of collection size on the fraction of relevant pages and subtopics with relevance.

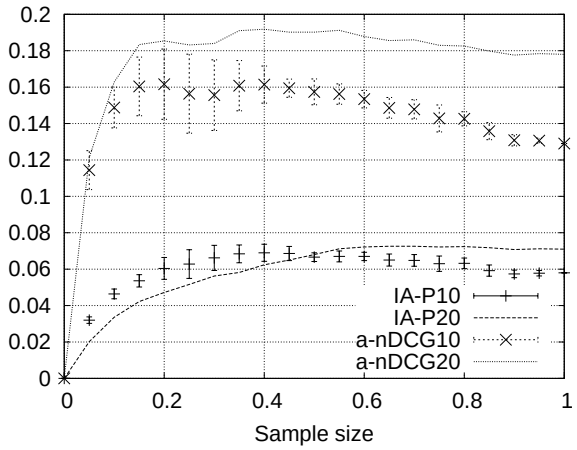

Figure 2: Impact of collection size on result diversity.

Table 1: Redundancy and diversity of relevant information

	Rel.docs/		# Subtopics Top 10		
Sample	**Topic**	**Subtopic**	**Inf.**	**Nav.**	**All**
10%	9.33	3.88	41.6	2.2	43.8
20%	16.41	5.75	47.4	2.0	49.4
30%	23.70	7.87	47.6	2.4	50.0
40%	31.37	9.90	47.8	2.6	50.4
50%	39.12	11.86	46.8	2.4	49.2
60%	46.32	13.73	46.2	1.6	47.8
70%	53.60	15.41	44.8	2.4	47.2
80%	61.46	17.17	43.4	2.4	45.8
90%	69.18	19.00	40.2	2.0	42.2
100%	76.71	20.88	39.0	2.0	41.0

3. REDUNDANCY AND DIVERSITY

We now analyse the effect of reducing the collection size on the redundancy of relevant information, and on the diversity or coverage of subtopics in the top of a retrieval run. Table 1 shows the number of relevant pages per topic and subtopic (columns 2 and 3). What is the impact of collection size on the redundancy of relevant information? The number of relevant pages and hence the redundancy steadily increases with the sample size. Eventually the collection contains many relevant document per topic and subtopic. Our second postulate also holds.

What is the impact on the diversity or coverage of subtopics in the top of the ranking? Table 1 shows the number of informational, navigational and total subtopics covered by a relevant document in the top 10 of our full-text run (columns 4, 5 and 6 respectively) when restricted to the sample. We see that initially the number of subtopics is increasing due to the increasing coverage for the smallest samples, but then peaks and tapers off due to the increasing redundancy for the larger samples.

Our analysis shows that collection size has a larger impact on redundancy than on the coverage of topics. This implies that the diversity at a fixed depth decreases with collection size, except for the smallest samples where coverage is still increasing noticeably.

4. RETRIEVAL EFFECTIVENESS

Finally, we analyse the effect of reducing the collection size on the performance on the TREC 2009 Web Track's Diversity Task's test collection. Figure 2 shows the impact of collection size on diversity performance. The top two lines show the α-nDCG measure at cutoffs 10 and 20, the bottom two show the IA-P measure at cutoffs 10 and 20. We show the variance over the 5 sample iterations only for the α-nDCG@10 and IA-P@10 to keep the figure readable. The variance is similar at rank 20. variance is largest between 20% and 30% of the collection. We see an initial increase of performance at sample sizes below 15% of the collection. After that, however, the performance doesn't increase further and remains relatively stable across sample sizes of 30% and above. In fact, the performance at rank 10 is actually decreasing. This is in line with the results in Table 1, supporting the validity of the measures.

Our analysis leads to the remarkable conclusion that when result diversity is of importance, we can improve performance by randomly removing more than two-thirds of the results from the collection or from a retrieval run.

5. CONCLUSIONS

We analysed the impact of collection size on relevance, coverage, redundancy and diversity. We found that the number of relevant documents increases, but the coverage of subtopics quickly saturates. As a result the redundancy of information steadily increases leading to a lower diversity of results. This leads to the remarkable conclusion that, when result diversity is of importance, we can improve performance by randomly removing the majority of the results—this will significantly reduce the redundancy and only marginally affect the subtopic coverage.

Our results are based on a standard full-text run—which does not do a very good job at retrieving diverse results—and an ideal diverse ranking would suffer from removing random results. However, it also makes a call to caution to any claim for a technique to diversify results. Any such techniques might improve in whole or in part due to an arbitrary reduction of the result-list.

In future research we investigate the impact of information redundancy, study better ways of reducing the collection than random sampling, and address the notion of an optimal collection size.

REFERENCES

[1] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview the TREC 2009 web track. In *The Eighteenth Text REtrieval Conference (TREC 2009) Notebook*. National Institute for Standards and Technology, 2009.

[2] D. Hawking and S. Robertson. On collection size and retrieval effectiveness. *Information Retrieval*, 6:99–150, 2003.