Bias and the Limits of Pooling

Chris Buckley
Sabir Research, Inc
cabuckley@sabir.com

Darrin Dimmick, Ian Soboroff,
Ellen Voorhees
National Institute of Standards and Technology
{darrin.dimmick, ian.soboroff,
ellen.voorhees} @nist.gov

ABSTRACT

Modern retrieval test collections are built through a process called pooling in which only a sample of the entire document set is judged for each topic. The idea behind pooling is to find enough relevant documents such that when unjudged documents are assumed to be nonrelevant the resulting judgment set is sufficiently complete and unbiased. As document sets grow larger, a constant-size pool represents an increasingly small percentage of the document set, and at some point the assumption of approximately complete judgments must become invalid. This paper demonstrates that the AQUAINT 2005 test collection exhibits bias caused by pools that were too shallow for the document set size despite having many diverse runs contribute to the pools. The existing judgment set favors relevant documents that contain topic title words even though relevant documents containing few topic title words are known to exist in the document set. The paper concludes with suggested modifications to traditional pooling and evaluation methodology that may allow very large reusable test collections to be built.

Categories and Subject Descriptors: H.3.4 [**Information Storage and Retrieval**]: *Systems and Software—Performance evaluation*

General Terms: Experimentation, Measurement

Keywords: retrieval test collections, pooling

1. TITLE WORD BIAS

The AQUAINT 2005 test collection was constructed as a joint product of the TREC 2005 HARD and robust tracks. The document set is the AQUAINT newswire collection, a set of more than 1 million newswire articles, and the topics are a set of 50 topics that had been used in previous TREC tasks with the TREC Disks4&5 document set. A diverse set of 50 runs were pooled to a depth of 55. Prior experience with pooling suggested this would be sufficient for an adequate test collection. Comparing evaluation results when runs are scored with and without the unique relevant documents of the current group [4] resulted in an average difference in MAP scores of 3.2% and a maximum difference of 23.1%. While not necessarily an indicator of a problem, a maximum difference of more than 20% is large, and this prompted a closer examination of the test collection.

The run with the greatest change in MAP score is run `sab05ror1`. This run is a routing run in which the existing Disks4&5 relevance judgments were used to create an optimal query based on Dynamic Feedback Optimization [2]. By design, the queries created for the run describe the relevant documents in the training set; as a result, the run put relatively less emphasis on terms appearing in the topic statement than did the other runs. Significantly, the relevant documents uniquely retrieved by this run also contain many fewer topic title words than the remaining relevant documents. We formalize this idea below.

Define the measure *titlestat* as the fraction of a set of documents, C, that a topic title word occurs in. For each word in the title of the current topic that is not a stop word, calculate the fraction of C that contains that word, normalized by the maximum possible fraction. (The normalization is necessary because a title word might have a collection frequency smaller than $|C|$.) Average over all title words for the topic, then average over all topics in the collection. A maximum value of 1.0 is obtained when all the documents in the set contain all topic title words; a minimum value of 0.0 means that all documents in the set contain no title words at all. Next, define the more specific measure *titlestat_rel* as titlestat computed over the set of relevant documents for each topic. The titlestat_rel value for the Disk4&5 collection is 0.588 and for the AQUAINT collection it is 0.719. The titlestat value computed over the unique relevant documents retrieved by the `sab05ror1` run is 0.530. The difference in titlestat_rel for the two collections is highly significant ($p = 6.25 \cdot 10^{-10}$ according to a paired t-test), and is very consistent across topics, with the AQUAINT collection having higher titlestat_rel for 48 of 50 topics.

While there are a number of differences between the Disks4&5 and AQUAINT document sets (time period covered, number of documents in the set, average document length, etc.), none is a plausible explanation for why the pattern of occurrence of topic title words should differ so dramatically in the relevant sets. Further, with the difference occurring for 48 of 50 topics, the difference is not attributable to characteristics of certain topic types such as those with many relevant documents. Instead, the pooling used in TREC 2005 produced a sample of the true judgment set for the AQUAINT collection that is biased against documents containing few topic title words. Future systems that do not have a similar bias in their retrieved sets would be evaluated unfairly by the collection, just as the `sab05ror1` would have been unfairly evaluated if not part of the pool.

The rationale for the too-shallow pools claim is a counting argument. Using a pool depth of λ, traditional pooling

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.
ACM 1-59593-369-7/06/0008.

adds the documents ranked 1–λ to the pool. Topic title words are intended to be highly descriptive of the material being sought, so retrieval systems purposely weight these terms highly. This results in documents containing many topic title words being ranked before documents containing fewer title words. Since in general the number of documents containing a given word will increase as the total number of documents increases, the number of documents containing topic title words will increase as the collection size increases. "Too shallow" means that λ is small relative to the number of documents in the collection: the absolute number of documents containing topic title words exhausts the available space in the pools to the exclusion of other types of documents.

2. TOWARD LARGE REUSABLE TEST COLLECTIONS

For the AQUAINT 2005 test collection we have the `sabo5ror1` run to demonstrate the existence of relevant documents that contain few topic title words. The collections built in the the TREC terabyte track do not have an equivalent "smoking gun" run, but doubts regarding the viability of traditional pooling for documents sets in the terabyte range seem justified given this counting argument. The title-stat_rel values for the TREC 2004 and TREC 2005 terabyte collections are 0.889 and 0.898. While using deeper pools may be in theory be feasible to produce an unbiased collection for the AQUAINT collection, it is clearly impossible to do for reasonable cost for these much larger collections. New approaches to building very large, reusable test collections are needed. Several possible approaches, with advantages and drawbacks, are briefly discussed here.

Engineering the topics. Construct topics such that title word occurrence is not a problem.

Drawback: applicability of any research results will then be limited to tasks where such narrow topics are required; the research results would not indicate anything about retrieval effectiveness in general.

Forming pools more efficiently. Build pools in ways designed to include documents from deeper in the rankings. Several methods have been shown to locate most relevant documents or to estimate conventional measures using a fraction of the currently judged documents; an assessment regime could apply these techniques within the current pooling "budget" and explore a much deeper pool. One such method that we have examined is move-to-front pooling [3]. An alternative method is random sampling, which can estimate MAP scores accurately with judgments from 10-20% of the traditional pool [1].

Drawbacks: How these methods interact with bias is unknown. The efficiency savings is likely insufficient to solve the problem.

Encouraging different retrieval approaches. Participants in a collection-building exercise could be required to perform other types of runs such as manual feedback runs, routing runs, or "query track"-style runs with combinations of multiple manual queries to enrich the pools in the way that manual runs have historically done.

Drawback: Run diversity is not a complete solution in itself: the AQUAINT collection was formed with many different kinds of runs which involved humans to a lesser or greater extent, and it is still biased. However, run diversity did allow the bias in the AQUAINT collection to be detected.

Engineering the judgment set. Continue traditional pooling, but then downsample the resulting judgments to a fair subset, discarding any judgments not in the subset. The major advantage of this approach over the others is that we can experiment with it given the current collections. Bpref can be used to evaluate the runs with the fair partial judgment set.

Drawback: It is currently unknown how to construct (or define) a fair sample.

Use of different evaluation measures. Preliminary examination of the effect bias has on evaluation measures shows that MAP and bpref are affected in different directions: runs that have a bias different from the judgment set tend to have their MAP scores underestimated and their bpref scores overestimated. If both measures show a statistically significant result in the same direction when comparing two runs, then it is very unlikely that the collection bias is the cause of the observed difference.

Drawback: The circumstances in which MAP and bpref differ needs to more fully explored. Researchers using the AQUAINT and larger collections must report both MAP and bpref.

3. CONCLUSION

Obtaining human judgments is the expensive part of building retrieval test collections. Pooling with a constant pool size fails as collection size grows in that the resulting judgment set becomes a biased sample of the complete judgment set and thus systems might not be fairly compared.

We present evidence that one type of bias, bias towards documents containing topic title words, exists in the 2005 AQUAINT collection. We suggest that given the state-of-the-art of current retrieval systems, such bias will exist in any very large test collection built using traditional document pooling techniques.

Even though biased collections can be used with care, it is much preferable to construct unbiased collections to begin with. Promising avenues to pursue to build very large reusable test collections include constructing pools using techniques designed to include documents from deeper in the systems' rankings and engineering small, but fair, judgment sets.

4. REFERENCES

[1] Javed A. Aslam, Virgiliu Pavlu, and Emine Yilmaz. A sampling technique for efficiently estimating measures of query retrieval performance using incomplete judgments. In *Proceedings of the 22nd ICML Workshop on Learning with Partially Classified Training Data*, pages 57–66, August 2005.

[2] Chris Buckley and Gerard Salton. Optimization of relevance feedback weights. In Proceedings of SIGIR 1995, pages 351–357, 1995.

[3] Gordon V. Cormack, Christopher R. Palmer, and Charles L.A. Clarke. Efficient construction of large test collections. In Proceesings of SIGIR 1998, pages 282–289, 1998.

[4] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In Proceedings of SIGIR 1998, pages 307–314, 1998.