Poison Pills: Harmful Relevant Documents in Feedback

Egidio Terra
Faculty of Computer Science
PUC/RS
Porto Alegre, Brazil
egidio@inf.pucrs.br

Robert Warren
School of Computer Science
University of Waterloo
Waterloo, Canada
rhwarren@uwaterloo.ca

Categories and Subject Descriptors

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

General Terms

Performance, experimentation

Keywords

relevance feedback, query expansion

1. INTRODUCTION

Relevance feedback (RF) aims to overcome query and document disagreements on vocabulary, and user and system disagreements on relevancy for a given information need. It can be implemented as an interactive (IRF) or automatic (pseudo-relevance feedback - PRF) process and relies on the initial document ranking returned by the system for the original query. The underlying assumption is that the initial retrieval will yield the relevant documents to refine the query[4, 7]. It is expected that they will provide new terms, such as synonyms, that will improve the original query.

The feedback performance of retrieval systems depends on many factors and a great variability in precision occurs from topic to topic [1]. On average, the overall system performance improves after feedback, but for some topics the average precision achieved is actually lower than the initial run, particularly in PRF.

This effect may be caused by the use of non-relevant documents (in the case of PRF) or due to the use of a single feedback parameter set for all topics [2]. In this work we present a new alternative explanation for lowered topic performance: under specific feedback parameter settings, relevant documents may act as "poison pills"[9] and harm topic precision after feedback.

2. FEEDBACK PARAMETERS

There are many possible parameter settings for relevance feedback. For instance, one may prefer use passages rather than documents. The optimal number of documents from the initial run considered is not known in advance neither the number of terms to be added to the query. Harman [4] discussed term re-weighting and the number of query terms. In the case of PRF the documents used for feedback are normally selected from the top-k ranked in the initial run but they could be the centroids of clusters from these top-k documents or the top-k interspaced documents. Shen and Zhai [8] discussed different strategies in the selection of the feedback documents. Billerbeck and Zobel [2] reviewed parameter setting issues in PRF.

The different retrieval models may need different strategies, such as Rocchio method [7] in the Vector Space Model and the divergence minimisation approach of language models for information retrieval [10]. Naturally, wrong choices will harm effectiveness and there is no definitive method to choose PRF parameters. The problems are similar in IRF except for relevant documents as this information is supplied by the user. However, since the number of documents that the user is able to read is limited, the document selection is constrained [3].

3. POISON PILLS: BAD DOCUMENTS FOR FEEDBACK

We used TREC 6, 7 and 8 *adhoc* track for evaluation, a total of 150 topics, along with the known relevant documents. We used the runs from four systems participating in the RIA workshop [5]: Clarit (TF-IDF), CMU (Lemur-language model), Sabir (SMART-Vector Space) and City (Okapi-probabilistic). They represent distinct information retrieval models as we try to examine the search space of parameters/alternatives in an ad hoc retrieval task. The feedback mechanisms used by these systems are also distinct among them.

For each of these systems we used a single relevant document per topic as the only feedback and, because of the relevance, we expected an increase in precision on the corresponding topic. Surprisingly, 299 documents used for feedback made the precision drop in all four systems and we refer to these documents as bad relevant documents.

Figure 1 depicts the number of topics that these documents affect; a total of 47 topics out of the 150 have one or more bad relevant document. The drop in performance is substantial, in some cases the average precision is 0.10 lower than the baseline. They represent a small fraction of all relevant documents; we plot the ratio of bad relevant documents per topic in Figure 2. Overall, 5.33% of the relevant documents performed poorly for single document feedback. This number is small thus the chances of selecting one of these documents is low. However, we also look at the rank distribution of these documents, as depicted in Table 1, where the number of bad and relevant documents are listed in different rank levels in the initial retrieval for the four systems. Some of the bad and relevant documents do not appear in the top-1000 retrieved documents of the four systems run. An inspection on the various depth levels show that bad

CIKM'05, October 31–November 5, 2005, Bremen, Germany.
ACM 1-59593-140-6/05/0010.

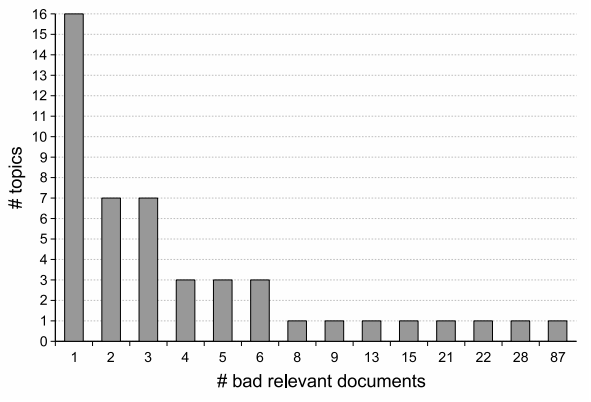

Figure 1: Histogram of the number of poison pills per topic

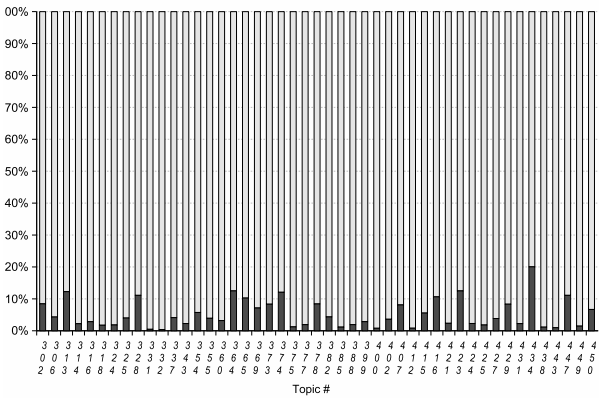

Figure 2: Fraction of bad rel documents per topic

Table 1: Rate of bad docs at various levels(4 systems)

Rank	# bad docs	# relevant docs	%
5	31	326	8.68
10	43	568	7.04
15	51	763	6.27
20	60	919	6.13
40	85	1381	5.80
100	148	2137	6.48
1000	269	3940	6.39
Overall	299	5315	5.33

documents occur in higher percentage in the early rankings.

4. DISCUSSION

Further analysis is necessary to understand the nature of these bad documents. In particular, we are attempting to characterise a bad document. Using one document for feedback is not usual in PRF, but it is not unrealistic in IRF. Some hypothesis may be explanatory to these poison pills:

• Document talks about several topics. In this case, the feedback mechanism may be selecting a term from another topic in the document. Used TREC documents are not normally multi-topic.

• Binary relevance judgements. Some documents may be marginally relevant, be on topic but with emphasis in another aspect of it [6].

• Feedback parameters not set for single docs. The number of documents and terms to be added to the query may need a different setting in the case of a single document. However, preliminary tests on multi-document feedback runs have indicated that while the problem is attenuated, the presence of bad documents will still lower the average precision.

• Feedback may need non-relevant documents. In general pseudo-relevance feedback does assume relevancy for top-k documents. Dunlop [3] further suggest that documents that do not match the query could be used.

• More feedback documents are needed. The feedback mechanisms may not be getting enough information to expand queries.

5. CONCLUSION AND FUTURE WORK

Feedback improves retrieval in average over many topics but not necessary for all topics individually. A possible reason for some topics to have a drop in performance is the existence of bad relevant documents. We have presented some numerical evidence that some relevant documents are harmful for feedback when used alone. Further analysis will be performed using more relevant documents for feedback, including one bad and other relevant documents.

6. REFERENCES

[1] N. Alemayehu. Analysis of performance variation using query expansion. *J. Am. Soc. Inf. Sci. Technol.*, 54(5):379–391, 2003.

[2] B. Billerbeck and J. Zobel. Questioning query expansion: an examination of behaviour and parameters. In *CRPIT '27: Proceedings of the 15th conference on Australasian database*, pages 69–76, 2004.

[3] M. D. Dunlop. The effect of accessing nonmatching documents on relevance feedback. *ACM Trans. Inf. Syst.*, 15(2), 1997.

[4] D. Harman. *Information Retrieval: Data Structures and Algorithms*, chapter Relevance feedback and other query modification techniques. Prentice-Hall Inc., 1992.

[5] D. Harman and C. Buckley. The NRRC Reliable Information Access (RIA) workshop. In *SIGIR '04*, 2004.

[6] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4), 2002.

[7] J. Rocchio. *The SMART Retrieval System: Experiments in Automatic Document Processing*, chapter Relevance feedback in information retrieval. Prentice-Hall Inc., 1971.

[8] X. Shen and C. Zhai. Active feedback - UIUC TREC-2003 HARD experiments. In *TREC-2003*.

[9] R. H. Warren and T. Liu. A review of relevance feedback experiments at the 2003 reliable information access (RIA) workshop. In *SIGIR '04*, 2004.

[10] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM '01*. ACM Press, 2001.