The Healing Power of Poison: Helpful Non-relevant Documents in Feedback

Mostafa Dehghani Samira Abnar Jaap Kamps
Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands
{dehghani,s.abnar,kamps}@uva.nl

ABSTRACT

The use of feedback information is an effective approach to address the vocabulary gap between a user's query and the relevant documents. It has been shown that some relevant documents act like "poison pills," i.e. they hurt the performance of feedback systems despite the fact that they are relevant. In this paper, we study the positive counterpart of this by investigating the helpfulness of non-relevant documents in feedback. In general, we find that although documents that are explicitly judged as non-relevant are normally assumed to be poisonous for feedback systems, sometimes considering high-scored non-relevant documents as a positive feedback helps to improve the performance of retrieval. In our experimental data, we observe a considerable fraction of non-relevant documents in higher ranked positions of the initial retrieval run, for most of the topics. Hence, by ignoring the potential value of non-relevant documents, we may loose a lot of useful information.

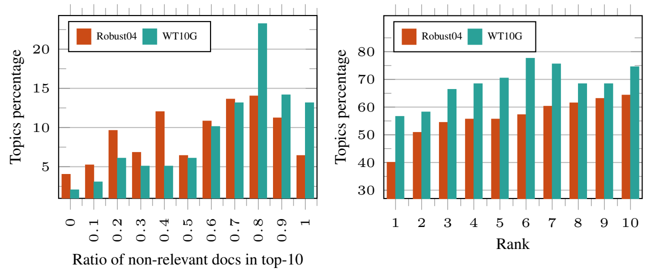
We investigate the potential contribution of non-relevant documents using existing state-of-the-art feedback methods. Our main findings are the following. First, we find that some of the non-relevant documents are *exclusively helpful*, they improve retrieval on their own, and others are *complementary helpful*, they lead to further improvement when added to a set of relevant documents. Second, we discover that, on average, exclusively helpful non-relevant documents have a higher contribution to the performance improvement, compared to the complementary ones. Third, we show that non-relevant documents in topics with poor average precision in the initial retrieval are more likely to help in the feedback.

Keywords: Relevance Feedback, Helpful Non-relevant Documents

1. INTRODUCTION

> *Often, the only difference between a medicine and a poison is the dose. Some substances are extremely toxic, and therefore, are primarily known as a poison. Yet, even poisons can have medicinal value.*
>
> `Paracelsus, Father of Toxicology`

Query expansion based on feedback information is one of the classic approaches for improving the performance of information retrieval systems, especially when the user information need are complex to express precisely in a few keywords. True Relevance Feedback (TRF) systems try to enrich the user query using a set of judged documents, that their relevance is assessed either explicitly by the user or implicitly inferred from the user behavior. However, this information is not always available. Alternatively, Pseudo Relevance Feedback (PRF) methods, also called blind relevance feedback, assumes that the top-ranked documents in the initial retrieved results are all relevant and use them for the feedback model.



(a) Percentage of topics with different ratio of non-relevant docs in top-10 results **(b)** Percentage of topics with non-relevant docs at different ranks

Figure 1: Prevalence of non-relevant documents in top ranked positions.

Normally feedback documents that are annotated as relevant are considered to be beneficial for feedback and feedback documents that are annotated as non-relevant are expected to be poisonous, i.e. they supposedly decrease the performance of the feedback systems if they are used as positive feedback. Based on this assumption, some of the TRF methods, use non-relevant documents as negative feedback [14, 15] and some PRF methods try to avoid using these documents. For example, some PRF methods attempt on detecting non-relevant documents in order for being robust against their noises [5, 6], or they manage to partially use their content in the feedback procedure, like some of their passages [8, 10]. Although PRF methods use non-relevant documents, they do not directly intend to take advantage of them as helpful documents. In other words, most of the time, removing non-relevant documents from the feedback set of PRF methods leads to a better performance. However, it has been shown that the assumption that all relevant documents improve the performance of feedback systems as positive feedback documents is not always true and sometimes even the relevant documents act like "poison pills" and decrease the performance [13]. As a counterpart fact, we speculate that non-relevant documents might sometimes be helpful as positive feedback [4]. Thus, in this paper, we are investigating the potential healing power of poisonous documents.

Do we really need to think of dealing with non-relevant documents when we are only taking top-scored documents into consideration for the feedback? Based on an analysis of standard test collections, the answer is: yes we do; because they are very prevalent in the top rank positions. Figure 1a depicts the percentage of topics with different ratio of non-relevant documents in top-10 results retrieved using the KL-Divergence model, in the two standard TREC

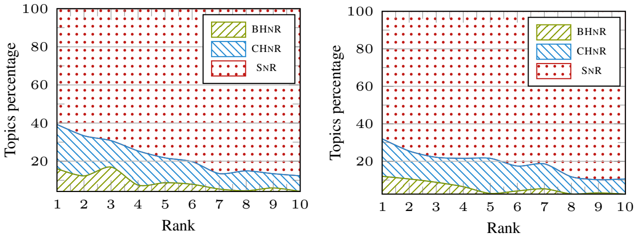
CIKM'16 , October 24 - 28, 2016, Indianapolis, IN, USA

ISBN 978-1-4503-4073-1/16/10. . . $15.00

DOI: http://dx.doi.org/10.1145/2983323.2983910

Table 1: Number of helpful non-relevant documents in different rank levels in the initial run.

level	Robust			WT10G		
	NR	BHNR	CHNR	NR	BHNR	CHNR
5	641	55 (%8.6)	106 (%16.5)	314	32 (%10.2)	45 (%14.3)
10	1404	82 (%5.8)	205 (%16.5)	671	58 (%8.6)	88 (%13.1)
15	2212	86 (%3.9)	219 (%9.9)	1050	66 (%6.3)	96 (%9.14)
20	3083	91 (%2.9)	232 (%7.5)	1423	71 (%5.0)	103 (%0.72)
50	8872	112 (%1.3)	241 (%2.7)	3862	74 (%1.9)	107 (%2.8)
100	18630	124 (%0.7)	250 (%1.3)	14040	79 (%0.6)	112 (%0.8)
500	66269	129 (%0.2)	257 (%0.4)	30827	84 (%0.3)	119 (%0.4)



(a) Robust04 dataset

(b) WT10G dataset

Figure 2: Percentage of topics with different types of non-relevant documents in different rank positions.

datasets: Robust04 and WT10G. It shows that for instance, about 30% of topics have seven or eight explicitly judged non-relevant document in their top-10 results for both datasets. Figure 1b also demonstrates the percentage of topics with non-relevant documents at different ranks. For instance, in WT10G, in more than 50% of topics, the top rank document is non-relevant or in more than 50% of topics, the document at rank two is non-relevant in both datasets. So, in general, there is a high probability of hitting a non-relevant document in top rank positions and ignoring their potential helpfulness is turning a blind eye to a lot of useful information.

We believe that every high-scored retrieved document, either judged as relevant or non-relevant, may contain information that can be a clue for understanding the complex information need of the user. Hence, if an ideal system is able to perfectly control the amount and the way each document contributes to the feedback model for each topic, not only it will not be hurt by a non-relevant document, but it will even be able to take advantage of its information to further improve the performance.

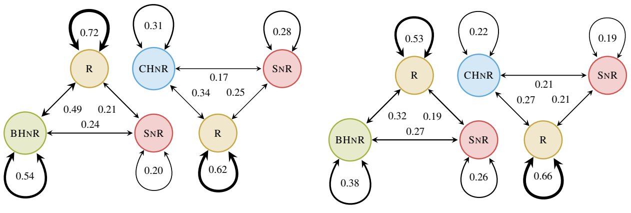
Generally, the main aim of this paper is to investigate the helpfulness of highly ranked non-relevant documents for improving further results by being used in the feedback methods. We break this down into the following research questions:

RQ1 *How can a non-relevant document help to improve retrieval performance?*

RQ2 *How large is the contribution of helpful non-relevant feedback documents?*

RQ3 *Does the helpfulness of the non-relevant documents depend on the quality of the initial retrieved results?*

For the sake of this study, we try to select the clearest and most explicit examples of helpful non-relevant documents (HNR), i.e. retrieved documents that are judged as non-relevant but considering them in the feedback set leads to improvement in the performance of retrieval, utilising the existing state-of-the-art feedback methods. Based on our observations, we divide HNR into two groups: Bridge Helpful Non-relevant (BHNR) documents that are able to improve the performance individually, and Complementary Helpful Non-relevant (CHNR) that further improve the performance if they are employed together with a set of relevant documents.



(a) Robust04

(b) WT10G

Figure 3: Intra and inter similarity of relevant (R), SNR, BHNR, and CHNR documents.

We use the Robust04 with 528,155 newswire documents and 249 topics, as well as WT10G with 1,692,096 web documents and 99 topics as the test collections, which are different in terms of both size and genre of documents. We use the KL-Divergence model, with Dirichlet smoothing as the retrieval model in all of the experiments. We employ three state-of-the-art feedback methods, relevance model (RM3) [1], MEDMM [11], and SWLM [2, 3]. It is important to note that we investigate only documents that are explicitly judged as non-relevant, and ignore unjudged documents.

2. HELPFUL NON-RELEVANT DOCUMENTS

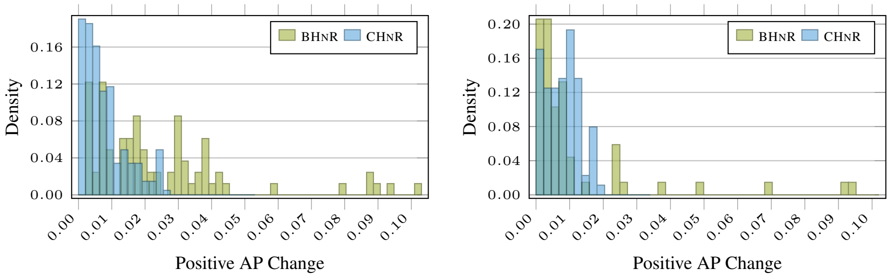
In this section, we address our first research question: "How can a non-relevant document help to improve retrieval performance?" Typically, it is expected that using a non-relevant document as positive feedback would cause a decrease in the average precision (AP). However, we show that some non-relevant documents not only do not hurt the feedback performance but they will even improve it.

Some hypothesis may be explanatory to this kind of documents. The first thing that comes to mind is that this is the effect of judgement noises and these documents are relevant but they have been misjudged. The second argument is that these documents are marginally relevant to the topic but with a focus on an aspect of the topic which is not related to the user information need or these documents are generally relevant to the topic but have no new valuable information to satisfy the user and because of using *binary* relevance judgments, they are annotated as non-relevant. In this case, they might contain terms that are helpful for query expansion. There is a discussion in linguistics that every utterance has two basic parts, *theme*, which indicates the topic of discourse, and *rheme*, which gives new information about the theme. In the retrieval problem, in rather crude terms, we may say that the theme of a document states what it is about, and the rheme of a document expresses what the searcher wishes to find in it [7]. So, for the searcher, relevance is a function of both theme and rheme of the document based on her current interests and his personal state of knowledge, while for the feedback method relevance is a function of the place of the document in the current state of knowledge as a whole. Thus, the mere annotation of a document as non-relevant for a topic does not imply that the document is not useful as positive feedback for that topic.

2.1 Bridge HNR

To understand the properties and behavior of HNR documents, we are attempting to extract the most explicit instances of HNR in our test collection and study their characteristics. To this end, for each of the feedback methods we use a single non-relevant document appearing in top-k retrieved documents per topic as the only feedback document and keep the track of change in AP of the corresponding topic. If the document makes the AP increase in all the systems, we mark that document as a HNR document.

We looked into several instances and found that regardless of the judgement noise, most of the times a non-relevant document improves the performance in single document feedback in two different situations: First, when the topic is a specific topic and the



(a) Robust04 dataset

(b) WT10G dataset

Figure 4: Distribution of HNR documents based on the amount of average change in the AP.

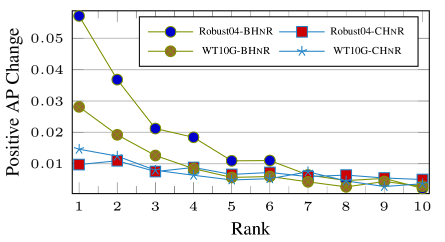


Figure 5: Average AP change caused by HNR documents in different rank positions.

HNR document is a broad topic document, which does not address directly the topic in a particular way, but some general frequent words in the language usage of the document help to retrieve more specific documents in the next feedback run. For example, in Robust04 dataset, topic 311 is “Industrial Espionage”, and one of the HNR documents is about “counter-espionage services for combating organised crime, terrorism and foreign operations”, and expanding query with generally related terms like terrorism, foreign, etc. boost the score of some relevant documents in the feedback run. Second, when the topic is a topic which the user is able to articulate it in different ways and the helpful non-relevant document does not address the user information need, but some of its frequent words express the topic in other ways. For instance, in Robust04 dataset,topic 348 is “creativity”, which based on the description of the topic, relevant documents are supposed to be about the definition of creativity, and one of the HNR documents is about “the founder of a company in Singapore, named Creative Technology”, which contains terms like “original thinkers”,“smart”, “innovation”, and these terms are related to the main concept of creativity. In both of these cases, the HNR document is not directly related to the topic, but it plays the role of a bridge to non-retrieved relevant documents. We refer to these documents as Bridge Helpful Non-relevant (BHNR) documents.

2.2 Complementary HNR

In a second experiment, we give one more chance to the remaining non-relevant documents to be helpful. To do this, for each topic, we check if adding the non-relevant document to the true relevance feedback set will further improve its performance. Thus, we examine the possibility of each individual non-relevant document to be helpful when it is used along with a set of relevant documents. We mark the documents that have a positive effect on the AP of the corresponding topic, in all the systems as a HNR document.

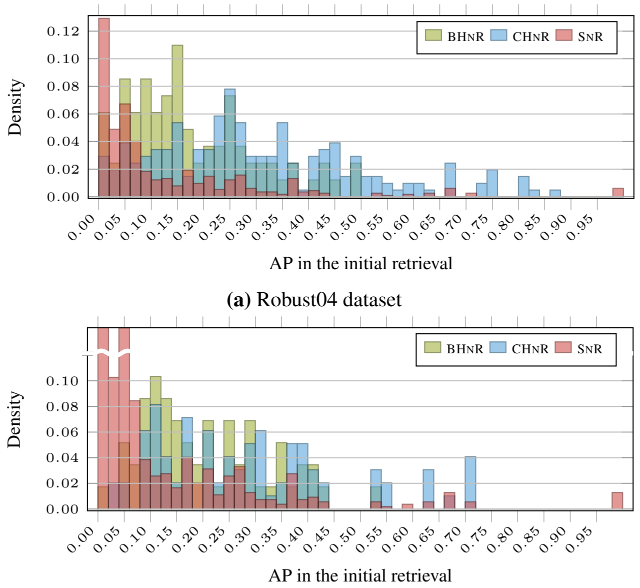
We explored different cases where a non-relevant document helps together with a set of relevant documents and discovered that this happens mostly in the situations where the HNR document is a multi-topic document, and technically it is partially relevant. This document is not able to be helpful in single-document feedback since its non-relevant part causes topic drift [6, 12]. However, together with other relevant documents, not only the effect of its non-relevant part is neutralised and compensated by the relevant documents, but its relevant part even reinforces the feedback models. We refer to these documents as Complementary Helpful Non-relevant (CHNR) documents. We examined the BHNR documents in our second experiment and observed that all of them are CHNR as well. So, we exclude them from CHNR documents to better analyse properties of each type. In Table 1, the rate of BHNR and CHNR documents are listed in different rank levels of the initial run in two datasets. As it is expected, in both datasets and in both types, HNR documents occur in a higher percentage in the early rankings. Moreover, an inspection of the various depth levels shows that from all HNR documents in top 500, a high percentage of them (60%-75%) take place in top-10. This might be one of the reasons that most of the pseudo relevance feedback methods work well with ten feedback documents [9]. We will restrict our analysis on the helpfulness of non-relevant documents to this level, i.e. up to rank ten. We also looked into the agreement of the three feedback methods in our experiments and observed that they behave quite differently. So, as a pragmatic choice, we study only the clear cases where all the methods agree on their helpfulness.

As it is pointed out in the introduction, we think that every high-scored retrieved document, either judged as relevant or non-relevant, can help feedback systems improve the performance. However, in this paper, we are not able to reveal the healing power of some of them, maybe because of the design of our experiments or due to the shortcoming of the existing feedback methods. Hence, we do not call them “bad” relevant document, and instead, we call them, Stubborn Non-relevant (SNR) documents. Figure 2 shows how likely a non-relevant document can be helpful in different top 10 rank positions. As expected, the percentages of the topics with HNR documents are higher in the higher ranks and generally, non-relevant documents are more likely to help along with a set of relevant documents, as CHNR, than individually as BHNR.

In order to understand the relation between HNR, SNR, and relevant documents, for each type of HNR documents, we select topics with at least two HNR, two SNR, and two relevant documents in their top ten results and calculate the intra and inter similarity of these groups of documents based on the average similarity of all document pairs, using JS-Divergence of their smoothed language models. Figure 3 shows the intra and inter similarity of relevant, SNR, BHNR, and CHNR documents in two datasets. As it is expected, in both datasets, the similarity of BHNR to the relevant documents and CHNR to the relevant documents is relatively higher than the similarity of SNR documents to the relevant documents. However, the intra similarity of BHNR documents is higher than CHNR documents. This observation is in accordance with the assumption that the BHNR documents are related to the topic (hence similar to the relevant documents) in general but they do not have the particular information which satisfies the searcher to be annotated as relevant. Also, CHNR documents are mostly multi-topic documents that have a part related to the query, but their non-relevant parts could be quite diverse, which leads to a low intra similarity.

3. IMPACT OF HNR DOCUMENTS

We now investigate our second research question: “How large is the contribution of helpful non-relevant feedback documents?” In order to evaluate the impact of HNR documents on the performance of feedback, we calculate the average of change of AP they make over all the employed feedback systems. For BHNR documents, we consider the amount of AP change compared to the initial run, and for CHNR, we consider the AP change compared to the feedback run without their presence in the feedback set. Figure 4 depicts the distribution of BHNR and CHNR documents based on the amount



(a) Robust04 dataset

(b) WT10G dataset

Figure 6: Distribution of different types of non-relevant documents based on the amount of AP in the initial retrieval.

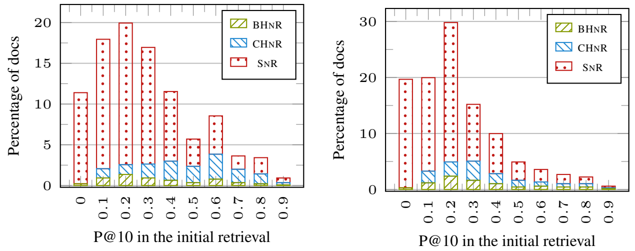
of average change in the AP in the two datasets, which shows that most of the HNR documents have a rather small contribution in AP. Figure 5 demonstrates the average AP change caused by the HNR documents in different rank positions. From the BHNR documents, in average, top rank documents have a greater contribution. However, based on the plots in this figure, the average amount of the positive contribution of the CHNR documents seems independent of the rank of the document. We have examined the contribution of all the BHNR documents in different ranks as CHNR and saw a roughly uniform plot as well. This reveals that since in CHNR experiments, involved relevant documents in the feedback can be from any rank, the feedback set is a mixture of documents from different ranks, hence the rank effect is disappeared.

4. IMPACT OF INITIAL PERFORMANCE

This section discusses our third research question: "Does the helpfulness of the non-relevant documents depend on the quality of the initial retrieved results?" In order to understand the relation between the quality of the initial retrieval and the helpfulness of the non-relevant documents, in Figure 6, we plot the distribution of BHNR, CHNR, and SNR documents based on the AP of the topic that these documents are retrieved for, in the initial run. Apparently, HNR documents occur mostly when the AP of the initial run is not too high. This is natural since when your baseline is not too good, it is easier to improve it, while when the performance of the baseline is high, it is easy to destroy it. In both datasets, CHNR documents are most likely to occur when both the AP of the initial run and the number of SNR in top-10 results are low. These properties are in accordance with the observations from Figure 7, where the percentages of BHNR, CHNR, and SNR in topics with different P@10 in the initial retrieval are presented. The more relevant documents exist in the feedback set, the greater the chance that a non-relevant document helps as a CHNR, and for low performing topics in terms of P@10, it is easier for a non-relevant document to improve the performance individually as a BHNR.

5. CONCLUSIONS

The main goal of this paper was to investigate the helpfulness of highly ranked non-relevant documents for improving further results by being used in the feedback methods. We designed some experiments based on the existing state-of-the-art feedback methods to assess the possibility of helpfulness of non-relevant documents. We found that some of the non-relevant documents can be helpful exclusively, and some others may help when they are employed as a complement to a set of relevant documents. We also discovered that in average compared to complementary ones, exclusively helpful non-relevant documents have a higher contribution to the improvement in performance . In addition, we showed that the non-relevant documents in topics with poor average precision in the initial retrieval are more likely to help in the feedback.



(a) Robust04 dataset

(b) WT10G dataset

Figure 7: Percentage of different types of non-relevant documents (out of all retrieved documents) in topics with different P@10 in initial retrieval.

This research is a primary step and further analysis is necessary to understand the nature of helpful non-relevant documents and how we can change feedback methods to be able to better take advantage of them. Also, as a direction from which this research is extendable, we are going to build a classifier to characterise the helpfulness of non-relevant documents based on the existing feedback methods.

Acknowledgments This research is funded in part by Netherlands Organization for Scientific Research through the *Exploratory Political Search* project (ExPoSe, NWO CI # 314.99.108), and by the Digging into Data Challenge through the *Digging Into Linked Parliamentary Data* project (DiLiPaD, NWO Digging into Data # 600.006.014).

REFERENCES

[1] N. Abdul-jaleel, J. Allan, W. B. Croft, O. Diaz, L. Larkey, X. Li, M. D. Smucker, and C. Wade. Umass at trec 2004: Novelty and hard. In *TREC-13*, 2004.

[2] M. Dehghani. Significant words representations of entities. In *SIGIR '16*, pages 1183–1183, 2016.

[3] M. Dehghani, H. Azarbonyad, J. Kamps, D. Hiemstra, and M. Marx. Luhn revisited: Significant words language models. In *CIKM '16*, 2016.

[4] M. D. Dunlop. The effect of accessing nonmatching documents on relevance feedback. *ACM Trans. Inf. Syst.*, 15(2):137–153, 1997.

[5] B. He and I. Ounis. Finding good feedback documents. In *CIKM '09*, pages 2011–2014, 2009.

[6] B. He and I. Ounis. Studying query expansion effectiveness. In *ECIR'09*, pages 611–619, 2009.

[7] W. J. Hutchins. On the problem of "aboutness" in document analysis. *Journal of informatics*, 1(1):17–35, 1977.

[8] X. Li and Z. Zhu. Enhancing relevance models with adaptive passage retrieval. In *ECIR'08*, pages 463–471, 2008.

[9] Y. Lv and C. Zhai. A comparative study of methods for estimating query language models with pseudo feedback. In *CIKM '09*, pages 1895–1898, 2009.

[10] Y. Lv and C. Zhai. Positional relevance model for pseudo-relevance feedback. In *SIGIR '10*, pages 579–586, 2010.

[11] Y. Lv and C. Zhai. Revisiting the divergence minimization feedback model. In *CIKM '14*, pages 1863–1866, 2014.

[12] C. Macdonald and I. Ounis. Expertise drift and query expansion in expert search. In *CIKM '07*, pages 341–350, 2007.

[13] E. Terra and R. Warren. Poison pills: Harmful relevant documents in feedback. In *CIKM '05*, pages 319–320, 2005.

[14] X. Wang, H. Fang, and C. Zhai. Improve retrieval accuracy for difficult queries using negative feedback. In *CIKM '07*, pages 991–994, 2007.

[15] X. Wang, H. Fang, and C. Zhai. A study of methods for negative relevance feedback. In *SIGIR '08*, pages 219–226, 2008.