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

Dehghani, M.; Abnar, S.; Kamps, J.

DOI
10.1145/2983323.2983910

Publication date
2016

Document Version
Final published version

Published in
CIKM'16

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/in-the-netherlands/you-share-we-take-care)

Link to publication

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
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

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 Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM’16, October 24 - 28, 2016, Indianapolis, IN, USA
© 2016 Copyright held by the owner/authors. Publication rights licensed to ACM. ISBN 978-1-4503-4073-1/16/10...$15.00
DOI: http://dx.doi.org/10.1145/2983323.2983910

Figure 1: Prevalence of non-relevant documents in top ranked positions. 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. 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 1(a) 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
Table 1: Number of helpful non-relevant documents in different rank levels in the initial run.

<table>
<thead>
<tr>
<th>level</th>
<th>Robust</th>
<th>WT10G</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NR</td>
<td>BHN R</td>
</tr>
<tr>
<td>5</td>
<td>641</td>
<td>55 (%8.6)</td>
</tr>
<tr>
<td>10</td>
<td>1404</td>
<td>82 (%5.8)</td>
</tr>
<tr>
<td>15</td>
<td>2212</td>
<td>86 (%5.9)</td>
</tr>
<tr>
<td>20</td>
<td>3083</td>
<td>91 (%2.9)</td>
</tr>
<tr>
<td>50</td>
<td>8872</td>
<td>112 (%1.3)</td>
</tr>
<tr>
<td>100</td>
<td>18630</td>
<td>124 (%0.7)</td>
</tr>
<tr>
<td>500</td>
<td>66269</td>
<td>129 (%0.2)</td>
</tr>
</tbody>
</table>

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

Figure 3: Intra and inter similarity of relevant (R), SnR, BHN R, and CHN R 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 [2]. 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
Robust04-CH

"original thinkers", "smart", "innovation", and these terms are related
we check if adding the non-relevant document to the true relevance
when the topic is a topic which the user is able to articulate it in dif-
words in the language usage of the document help to retrieve more
documents in top 500, a high percentage of them (60%-75%) take
inspection of the various depth levels shows that from all H
these documents as Complementary Helpful Non-relevant (CH
its relevant part even reinforces the feedback models. We refer to
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 H
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 5 shows the
distribution of BHNR and CHNR documents based on the amount
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.

2.2 Complementary HNR
In a second experiment, we give one more chance to the remain-
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 exam-
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

![Figure 4: Distribution of HNR documents based on the amount of average change in the AP.](image)

![Figure 5: Average AP change caused by HNR documents in different rank positions.](image)
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 re-