How Relevance Feedback is Framed Affects User Experience, but not Behaviour

Tripathi, Dhruv

ACM
2019

Tripathi , D , Medlar , A & Glowacka , D 2019 , How Relevance Feedback is Framed Affects User Experience, but not Behaviour . in CHIIR'19 : Proceedings of the 2019 Conference on Human Information Interaction and Retrieval . ACM , New York, NY , pp. 307-311 , ACM SIGIR Conference on Human Information Interaction and Retrieval , Glasgow , United Kingdom , 10/03/2019 . https://doi.org/10.1145/3295750.3298957

http://hdl.handle.net/10138/313971
https://doi.org/10.1145/3295750.3298957

How Relevance Feedback is Framed Affects User Experience, but not Behaviour

Dhruv Tripathi
University of Edinburgh
s1756548@sms.ed.ac.uk

Alan Medlar
University of Helsinki
alan.j.medlar@helsinki.fi

Dorota Głowacka
University of Helsinki
glowacka@cs.helsinki.fi

ABSTRACT

Retrieval systems based on machine learning require both positive and negative examples to perform inference, which is usually obtained through relevance feedback. Unfortunately, explicit negative relevance feedback is thought to have poor user experience. Instead, systems typically rely on implicit negative feedback. In this study, we confirm that, in the case of binary relevance feedback, users prefer giving positive feedback (and implicit negative feedback) over negative feedback (and implicit positive feedback). These two feedback mechanisms are functionally equivalent, capturing the same information from the user, but differ in how they are framed. Despite users' preference for positive feedback, there were no significant differences in behaviour. As users were not shown how feedback influenced search results, we hypothesise that previously reported results could, at least in part, be due to cognitive biases related to user perception of negative feedback.

CCS CONCEPTS

• **Information systems → Relevance assessment**;

KEYWORDS

relevance feedback, negative relevance feedback, user studies, experimental design, scientific literature search

ACM Reference Format:
Dhruv Tripathi, Alan Medlar, and Dorota Głowacka. 2019. How Relevance Feedback is Framed Affects, User Experience, but not Behaviour. In *2019 Conference on Human Information Interaction and Retrieval (CHIIR '19), March 10–14, 2019, Glasgow, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3295750.3298957

1 BACKGROUND

Traditional information retrieval (IR) systems present users with search results ranked by their relevance to a search query. IR systems tend to assume that users have specific search goals in mind and sufficient domain knowledge to formulate appropriate queries. In many situations, however, users have difficulty expressing their information needs [6]. In exploratory search, for example, users are characterised as attempting to learn or discover new information and may struggle to achieve their search goals without additional support [1, 2, 16, 17].

To support users, IR systems can use feedback mechanisms to supplement the traditional query interface. Relevance feedback (RF) is a mechanism by which users flag search results that are relevant to the current search [9, 13, 14]. By doing so, users can refine the scope of their search without explicitly describing what information they are seeking. Instead, the IR system infers which documents would satisfy the user based on those documents previously identified as relevant [12].

Traditionally, RF was used for query expansion: identifying new terms that could be included or reweighted in the search query [5]. Query expansion can be achieved by finding terms that are over-represented in relevant documents compared to background term frequencies in the corpus. However, it was demonstrated early on that using additional information from non-relevant documents improved performance [11]. What constitutes non-relevance differs from system to system. Non-relevant documents can be those that were *a*) assessed by the user and explicitly marked as non-relevant (explicit negative feedback) or *b*) assumed to be non-relevant because they were not marked as relevant (implicit negative feedback). In the latter case, there is no guarantee that users have assessed, or even seen, those documents assumed to be non-relevant.

In this study we are interested in non-relevant documents and negative feedback for two main reasons. First, retrieval systems are increasingly based on machine learning algorithms that require both positive and negative examples to perform inference (e.g. [3, 15, 18, 22]). Indeed, such systems assume that RF categorizes documents into relevant and non-relevant sets, whether users are doing so explicitly or not [8, 10]. Second, RF-based systems typically require an initial search query before any relevance judgments can be made. If we assume the query is at least broadly related to the user's information needs and that relevant documents are more likely to be ranked highly, then negative feedback should provide more information to the system than positive feedback [7, 12]. Indeed, we would argue that IR systems should be designed to optimise the quality of negative feedback.

It is therefore unfortunate that negative relevance feedback (NRF) - the explicit flagging of non-relevant documents - is thought to have a poor user experience in comparison to RF. Belkin *et al.* report that users worry about the unintended consequences of using NRF. That negative assessments could erroneously lower the ranking of relevant documents [4]. Ruthven and Lalmas suggest that this concern could be because the potential "harm" caused by negative feedback is unknowable: relevant documents that are missed out on are not presented to the user [19]. The effect of improper positive assessments, however, can be observed and corrected in subsequent iterations. Further criticisms of NRF include higher cognitive load,

CHIIR '19, March 10–14, 2019, Glasgow, United Kingdom

ACM ISBN 978-1-4503-6025-8/19/03…$15.00
https://doi.org/10.1145/3295750.3298957

unfamiliarity and general dislike [4, 13]. Indeed, when users have the opportunity to use NRF, it is either underutilised or not used at all [4, 21].

It is unclear, however, whether user experience of NRF is genuinely negative. The suggestion that the outcome of NRF is less observable than PRF is inaccurate. Any feedback, whether positive or negative, has an impact on which search results are shown to the user. As there is only a finite number of search results per page, there will always be results that do not get ranked sufficiently high and the "harm" of not seeing them is unknown in both cases. Other problems with NRF could merely be side-effects of user study or system/interface design. For example, while NRF is found to be more cognitively demanding, NRF is often implemented in addition to RF (users can give both positive and negative feedback). With respect to underutilisation of NRF, Belkin *et al.* [4] observed that some participants felt under time pressure to complete the experiment and consequently were less inclined to use NRF because it was unfamiliar. In real life, users have time to become comfortable with new interfaces, but in experimental settings may only get a single opportunity to perform their search task. Furthermore, past work does not make a distinction between the *mechanism* of negative assessment and its *implementation* in IR systems, where user experience can be impacted by other factors (interface, ranking algorithm, etc.). Indeed, Dunlop showed that NRF implementation is often inconsistent with our logical expectations of how it should operate, suggesting that our understanding of the user experience for NRF might be based on broken implementations [7].

In this study, users performed a categorisation task: they had to flag search results as either *relevant* or *not relevant* to a search query provided as part of the experiment. Participants were told that flagged documents could only be relevant (not relevant) and that unflagged documents were *de facto* not relevant (relevant). The experiments were designed to ensure there would be relevant, not relevant and ambiguous search results for each search query. Crucially, participants are performing the exact same task whether giving positive or negative feedback: categorising search results into two distinct sets. The only thing that changes is the set (relevant or not relevant) that flagging identifies. We show that how the feedback mechanism is framed (explicit positive/implicit negative versus implicit positive/explicit negative) has a significant impact on user experience, however, user behaviour and which documents are identified as relevant are indistinguishable from one another.

2 USER STUDY

2.1 Search System

Users were placed in a scientific literature search scenario, where documents are displayed in a list-like interface and relevance feedback is given per document. The interface to the system is shown in Figure 1. Each search task is limited to a single results page (users never see how their feedback influences search results in subsequent iterations). Instructions are displayed at the top of the page. Documents are presented in a list with a button next to each document for relevance feedback. Each search task shows 10 documents with 2-3 documents visible without scrolling (dependent on abstract length). A brief questionnaire at the bottom of the page was used to record immediate impressions of the current task. Search

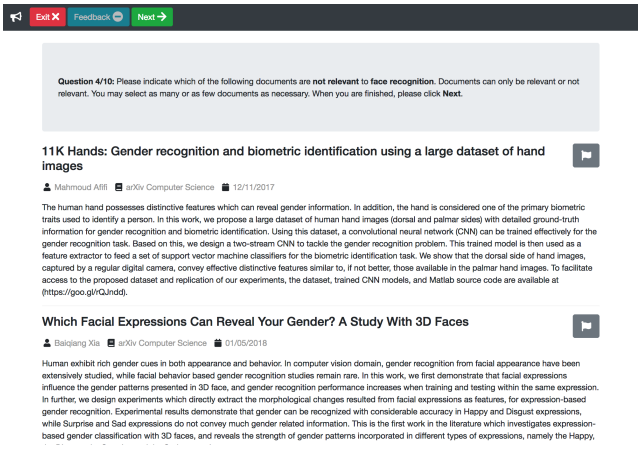


Figure 1: Screenshot of the search interface. Instructions for the search task are displayed at the top of the page. Documents can be flagged by clicking the icon in the right-hand margin. The Next icon takes the user to the next search task.

results were retrieved from a database of 170,367 Computer Science articles from arXiv (downloaded June 2018) using Okapi BM25 [20].

2.2 Study Design

We designed a within-subjects study where participants performed 10 search tasks. Each task required participants to either flag documents that were relevant to the stated search query (RF) or documents that were not relevant (NRF). An expert researcher from the machine learning domain designed five search queries well covered by our dataset. While the displayed documents were retrieved using these queries, participants were shown a different, related query (Table 1). For example, when presented with documents retrieved using the query "gender recognition", participants were asked to mark documents related to "face recognition". This was to ensure that search results contained a mixture of relevant, non-relevant and ambiguous documents.

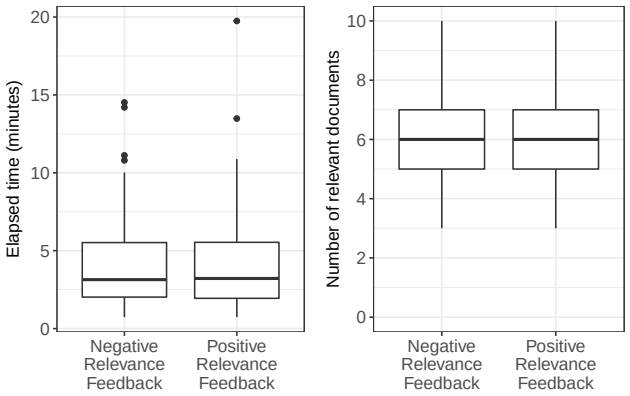
Participants performed assessments for each query twice: once to provide positive relevance feedback and once to provide negative relevance feedback. They were told explicitly that documents could only be relevant or not relevant. We included instructions at the top of each page using the same template for each search task:

> *"Question 1/10: Please indicate which of the following documents are **relevant/not relevant** to **search query**. Documents can only be relevant or not relevant. You may select as many or as few documents as necessary. When you are finished, please click Next."*

To avoid order effects, the 5 search queries were randomised and then repeated in the same order. This ensured queries were spaced out as much as possible: the first query being repeated as the sixth query, the second query as the seventh, etc. Whether the first search task was to mark documents as relevant or not relevant was randomised and then alternated for the remainder of the experiment. Finally, the order of the top ranking search results shown to the user were randomised.

True search query	Stated search query
gender recognition	face recognition
sentiment analysis	twitter sentiment analysis
wireless sensor networks	wireless energy efficiency
fake news twitter	fake news detection
deep learning medicine	clinical applications of deep learning

Table 1: List of search queries together with what participants were told the queries were. Stated queries differed to ensure search results were a mixture of relevant, ambiguous and irrelevant documents.



(a) Elapsed time **(b) Relevant documents**

Figure 2: Boxplots showing interaction data collected from search tasks using positive and negative relevance feedback. Subplot a shows the number of relevant documents is independent of feedback mechanism. Subplot b shows the same for the time taken to complete each search task.

2.3 Measures and Procedure

Prior to each experiment, users were shown how to complete search tasks via an instructional video. Participants were not informed that they would repeat each query with a different relevance feedback mechanism in the second half of the experiment.

Studies were conducted in a soundproof room in order to avoid distractions. The experiment was performed on a MacBook Air laptop with a 13 inch screen. There were no time limits to complete the experiments. At the beginning of the study, participants completed a consent form and a short questionnaire about their background. Participants then performed the 10 search tasks. Each search task contained the same instructional text (stated previously) at the top of the page and a short intermediate questionnaire at the bottom. The intermediate questionnaire asked users to assess their knowledge of the stated query and their level of comfort giving positive or negative relevance feedback depending on the current search task. After the final task users answered a questionnaire related to their overall assessment of the search tasks and different methods of relevance feedback.

To conclude the study, we conducted a semi-structured interview inquiring about each participants personal experience during the experiment. Each experiment lasted on average ~50 minutes in total and each participant was compensated with £10.

3 RESULTS

We recruited 24 participants (16 male and 8 female) between the ages of 18 and 35. The male to female ratio of 2:1 reflects the sex distribution in Computer Science in our institution. A majority of participants were finishing MSc students (22) in addition to 2 PhD students. All participants were proficient in English and had knowledge of artificial intelligence and machine learning. All participants reported experience with scientific literature search engines. Four participants were excluded from further study: 2 were pilot experiments and a further 2, based on post-experimental interviews, appeared to have misunderstood the instructions.

3.1 Negative Relevance Feedback Feels Unnatural

Users were asked to rate how natural/comfortable they found using positive or negative relevance feedback on a 5 point Likert scale at the end of each search task. We used ordinal logistic regression to understand whether there was a statistically significant difference between feedback mechanisms (ordinal R package, version 2015.6.28). Ordinal logistic regression is similar to multinomial logistic regression with the exception that the response variable is ordered, as in Likert scale responses. We controlled for the search query and accounted for repeated measures by including the participant as a random effect in the model. We performed a likelihood ratio test (LRT) comparing the full model with a reduced model that did not include a feedback mechanism term. There was a significant difference between positive and negative relevance feedback mechanisms based on users' perception of how natural or comfortable it felt ($LRT = 5.399; p = 0.02$). Users preferred giving positive over negative relevance feedback by a factor of 2.05 (95% CI [1.11, 3.78]). This finding was confirmed in the post-experiment questionnaire, where a majority of users (14/20) stated explicitly that they preferred giving positive over negative relevance feedback (14 positive relevance feedback, 4 negative relevance feedback, 2 no preference).

One possible explanation for positive relevance feedback feeling more natural is that negative relevance feedback is more cognitively demanding. However, only a slim majority of participants (12/20) felt this way. Seven participants felt the opposite, stating that positive relevance feedback was more demanding and 1 participant said that both modes felt the same.

3.2 Negative Relevance Feedback Feels Slower

A slim majority of participants (12/20) stated in the post-experiment questionnaire that performing positive relevance feedback felt faster than giving negative relevance feedback (5 thought negative relevance feedback faster, 3 thought neither was faster).

Figure 2a shows boxplots for the time taken to complete each search task. Users giving positive relevance feedback took on average 4.2 minutes per search task ($SD = 3.1$), whereas those giving negative relevance feedback took an average of 4.0 minutes

($SD = 2.7$). Repeated measures ANOVA found no significant difference between the time taken to perform the search task and the feedback mechanism ($F(1, 179) = 0.31; p = 0.578$). Indeed, positive relevance feedback was not significantly faster for the 12 people that said it was ($F(1, 107) = 0.006; p = 0.941$).

3.3 Users Disagree on Relevance, but are Highly Self-consistent

Whether using positive or negative feedback, participants identified the same number of relevant documents. Figure 2b shows boxplots for the number of relevant documents identified using both positive and negative relevance feedback. Users identified on average 5.87 relevant documents ($SD = 1.552$) using positive relevance feedback, and an average of 6.02 relevant documents ($SD = 1.58$) using negative relevance feedback. Repeated measures ANOVA found no significant difference between the number of relevant documents and the feedback mechanism ($F(1, 179) = 0.648; p = 0.422$).

In general, users identified the same documents as relevant whether using positive or negative feedback. Different users, however, thought different documents were relevant. Figure 3 is a graphical representation of document relevance across all experiments. Each search task is shown as a grid of documents (rows) and participants (columns). For each document we identified whether participants' feedback was consistent between positive and negative feedback, i.e. whether documents that received positive feedback did not receive negative feedback. If documents were identified as relevant to the query with both positive and negative relevance feedback, it is coloured purple. If neither method found the document to be relevant, it is coloured white. If it was relevant only with positive feedback, then it is red, and only with negative feedback, blue. Rows are sorted by inter-participant agreement (users agreeing on document relevance) and columns are sorted by intra-participant agreement (consistency of feedback for a single participant).

Users are highly self-consistent. Irrespective of whether they use positive or negative relevance feedback, users tend to come to the same relevance judgments. Indeed, 89.7% of squares in Figure 3 are coloured either purple (relevant with both feedback mechanisms) or white (relevant in neither). Despite participants' feedback being self-consistent, relevance feedback between participants tended to disagree depending on search task. We therefore believe that users made genuine relevance judgments and were not simply looking for matching keywords in the snippet text for each document.

3.4 Qualitative Feedback Reveals Mixed Feelings

The post-experimental interviews revealed a complex picture of how participants felt about different methods of relevance feedback. Several participants reported that negative relevance feedback made intuitive sense: "No problems using negative feedback ... We can save time by removing documents which are completely non-relevant" [Participant 9]. Others, however, did not like giving negative feedback. One participant said "it is a bit weird and unnatural. The idea of flagging is to get things which are relevant because when you search for something you want relevant results" [P11]. Another participant commented that negative feedback was "more taxing because it required a deeper understanding of the topics"

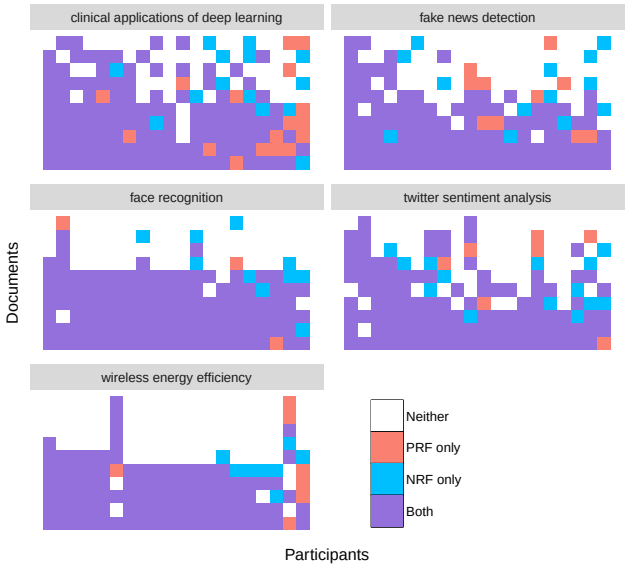


Figure 3: Visualisation of relevance feedback consistency. Each search query is shown as a grid of documents (rows) and participants (columns). Purple and white squares show documents where positive and negative feedback agree.

[P8]. Almost half of respondents, however, reported that after a few search tasks they felt more comfortable giving negative feedback despite their initial misgivings.

Irrespective of feedback mechanism, majority of participants claimed to have difficulties marking ambiguous documents and split evenly on whether ambiguous documents were better reported as relevant or not relevant. In the case of ambiguous documents, 75% of participants said they made the relevance judgments based solely on the search query. We assume that this strategy relates to the semantic meaning of search query terms and not exact keyword matches.

4 DISCUSSION AND CONCLUSIONS

While users tend to prefer positive relevance feedback over negative feedback, they performed consistently with both methods in terms of finding relevant documents and the time taken. First, implicit negative feedback appears to give the same results as explicit negative feedback, validating an assumption made by machine learning-based systems. Second, while giving negative feedback had a worse user experience than positive feedback, in our experiments it was exactly the same categorisation task, just framed in a different manner. Neither feedback mechanism could be objectively better or worse than the other, which is supported by users' behaviour. Therefore, any differences in experience can only be explained as the effect of cognitive biases. Retrieval systems might be able to improve user experience of negative feedback by presenting it using an intuitive analogy, e.g. pruning away bad things, to overcome these biases. As most systems only use positive feedback, this could lead to novel search systems. To conclude, negative feedback may have a worse user experience, but it is all in our heads.

REFERENCES

[1] Kumaripaba Athukorala, Dorota Głowacka, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. 2016. Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks. *Journal of the Association for Information Science and Technology* 67, 11 (2016), 2635–2651.

[2] Kumaripaba Athukorala, Alan Medlar, Antti Oulasvirta, Giulio Jacucci, and Dorota Glowacka. 2016. Beyond Relevance: Adapting Exploration/Exploitation in Information Retrieval. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. ACM, 359–369.

[3] Khadidja Belattar and Sihem Mostefai. 2013. CBIR using relevance feedback: comparative analysis and major challenges. In *Computer Science and Information Technology (CSIT), 2013 5th International Conference on*. IEEE, 317–325.

[4] Nicholas J Belkin, J Perez Carballo, S Lin, SY Park, SY Rich, P Savage, C Sikora, H Xie, C Cool, and James Allan. 1998. Rutgers' TREC-6 interactive track experience. In *TREC*. 221–229.

[5] Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)* 44, 1 (2012), 1.

[6] Sudatta Chowdhury, Forbes Gibb, and Monica Landoni. 2011. Uncertainty in information seeking and retrieval: A study in an academic environment. *Information Processing & Management* 47, 2 (2011), 157–175.

[7] Mark D Dunlop. 1997. The effect of accessing nonmatching documents on relevance feedback. *ACM Transactions on Information Systems (TOIS)* 15, 2 (1997), 137–153.

[8] Dorota Glowacka and John Shawe-Taylor. 2010. Content-based image retrieval with multinomial relevance feedback. In *Proceedings of 2nd Asian Conference on Machine Learning*. 111–125.

[9] Donna Harman. 1992. Relevance feedback revisited. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1–10.

[10] Sayantan Hore, Lasse Tyrvainen, Joel Pyykko, and Dorota Glowacka. 2014. A reinforcement learning approach to query-less image retrieval. In *International Workshop on Symbiotic Interaction*. Springer, 121–126.

[11] Eleanor Ide. 1971. New experiments in relevance feedback. *The SMART retrieval system: Experiments in automatic document processing* (1971), 337–354.

[12] Maryam Karimzadehgan and ChengXiang Zhai. 2010. Exploration-exploitation tradeoff in interactive relevance feedback. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 1397–1400.

[13] Diane Kelly and Xin Fu. 2006. Elicitation of term relevance feedback: an investigation of term source and context. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 453–460.

[14] Diane Kelly and Jaime Teevan. 2003. Implicit feedback for inferring user preference: a bibliography. In *Acm Sigir Forum*, Vol. 37. ACM, 18–28.

[15] Mohammed Lamine Kherfi and Djemel Ziou. 2006. Relevance feedback for CBIR: a new approach based on probabilistic feature weighting with positive and negative examples. *IEEE Transactions on Image Processing* 15, 4 (2006), 1017–1030.

[16] G. Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[17] Alan Medlar and Dorota Glowacka. 2018. How Consistent is Relevance Feedback in Exploratory Search?. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 1615–1618.

[18] A. Medlar, K. Ilves, P. Wang, W. Buntine, and D. Glowacka. 2016. PULP: A System for Exploratory Search of Scientific Literature. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1133 – 1136.

[19] Ian Ruthven and Mounia Lalmas. 2003. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review* 18, 2 (2003), 95–145.

[20] K. Sparck Jones, S. Walker, and S. E. Robertson. 2000. A probabilistic model of information retrieval: Development and comparative experiments. *Info. Proc. & Manag.* 36, 6 (2000), 779–840.

[21] Amanda Spink, Dietmar Wolfram, Major BJ Jansen, and Tefko Saracevic. 2001. Searching the web: The public and their queries. *Journal of the American society for information science and technology* 52, 3 (2001), 226–234.

[22] Xiang Sean Zhou and Thomas S Huang. 2003. Relevance feedback in image retrieval: A comprehensive review. *Multimedia systems* 8, 6 (2003), 536–544.