How Consistent is Relevance Feedback in Exploratory Search?

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

Dorota Glowacka
School of Informatics
University of Edinburgh
dorota.glowacka@ed.ac.uk

ABSTRACT

Search activities involving knowledge acquisition, investigation and synthesis are collectively known as exploratory search. Exploratory search is challenging for users, who may be unable to formulate search queries, have ill-defined search goals or may even struggle to understand search results. To ameliorate these difficulties, reinforcement learning-based information retrieval systems were developed to provide adaptive support to users. Reinforcement learning is used to build a model of user intent based on relevance feedback provided by the user. But how reliable is relevance feedback in this context? To answer this question, we developed a novel permutation-based metric for scoring the consistency of relevance feedback. We used this metric to perform a retrospective analysis of interaction data from lookup and exploratory search experiments. Our analysis shows that for lookup search relevance judgments are highly consistent, supporting previous findings that relevance feedback improves retrieval performance. For exploratory search, however, the distribution of consistency scores shows considerable inconsistency.

CCS CONCEPTS

• **Information systems** → **Personalization**; *Information retrieval diversity*;

KEYWORDS

relevance feedback, exploratory search, lookup search

ACM Reference Format:
Alan Medlar and Dorota Glowacka. 2018. How Consistent is Relevance Feedback in Exploratory Search?. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3269206.3269297

1 INTRODUCTION

Search activities are broadly categorised into lookup and exploratory search tasks [9]. In exploratory search, users are unfamiliar with a given search domain and want to either learn about, or investigate, a specific topic or data set. Users' lack of knowledge makes exploratory search challenging [15]: users have difficulty formulating search queries and have open-ended, or ill-defined, search goals that may change throughout a search session [2]. This is in contrast to lookup search, where users are performing fact retrieval or question answering. Users performing lookup search are assumed to have sufficient domain knowledge to formulate search queries and interpret results correctly.

In lookup search, relevance feedback has been shown to improve retrieval performance by enabling the search system to train a user model or help reformulate the search query [13]. In exploratory search, however, the benefits of relevance feedback have not been so clearly demonstrated (with the exception of content-based image retrieval, though this is not explicitly described as exploratory [16]). Nevertheless, recent work exploring the application of reinforcement learning in exploratory search relies on relevance feedback to provide the learning algorithm with positive and negative examples throughout a given search session [7, 10, 11]. These systems make the implicit assumption that users are capable of providing sufficiently consistent feedback from which to base learning, but this assumption remains unproven.

To quantify the quality of relevance feedback in interactive information retrieval, we created a permutation-based scoring metric to assess feedback inconsistency over a search session. Our approach estimates weights for feature importance from document-level relevance feedback, which is used to score feedback from the previous search iteration. Lower scores suggest that relevance feedback was congruent between consecutive iterations, whereas higher scores quantify the level of inconsistency. This work presents the following contributions: *1*) a novel permutation-based metric for relevance feedback inconsistency, *2*) a *post hoc* analysis of user interaction data from a relevance feedback-based search engine. We demonstrate that there are significant differences in feedback inconsistency between lookup and exploratory search tasks.

2 RELATED WORK

Over the last couple of decades, there have been numerous studies of user behaviour in exploratory search [9]. In general, users start with a broad query or concept formulation, which gradually narrows as the search progresses. This concept narrowing in exploratory search also affects users' click behaviour. For example, users click more documents at the beginning of a search session and then become more selective as the search progresses and their search intent becomes more pronounced [15]. Users tend to click on more results if they are presented with documents commensurate with their level of knowledge, i.e. beginners are more likely to click on overviews, whereas experts prefer more detailed results [4]. Taking user feedback into consideration and thus contextualising the results also increases the number of clicks, particularly in exploratory search, e.g. after clicking on an article about apples (fruit) as opposed to computers, the user is presented with more

CIKM '18, October 22–26, 2018, Torino, Italy

ACM ISBN 978-1-4503-6014-2/18/10.
https://doi.org/10.1145/3269206.3269297

search results related to the health benefits of apples, which results in higher user engagement [8].

Studies of search behaviour in exploratory search led to the creation of models of exploratory information seeking, the most notable being the Information Foraging Theory [12]. The key idea is that users decide what results to click according to the expectation of information gain. Berry-picking is another human-centered model that assumes search is a constantly evolving phenomenon with the user updating their cognitive model of information need [5]. Most of the existing models of exploratory search behaviour rely heavily on implicit and explicit user feedback, in particular, the click data. Although click models have been well-studied in the context of lookup search [6], little is understood in terms of how reliable and consistent click data is in exploratory search [1, 4].

3 METHODS

3.1 Approach

We assume the existence of an interactive information retrieval system where in a given search iteration, t, users are presented with a set of documents, D^t, for which they can provide binary (positive or negative) relevance feedback. The search system uses the relevance feedback to decide the next set of documents, D^{t+1}, to present to the user. Relevance feedback categorises D^t into two sets: P^t and N^t – containing documents that received positive feedback and negative feedback, respectively, where $P^t \cup N^t = D^t$ and $P^t \cap N^t = \emptyset$. We consider explicit negative feedback and implicit negative feedback (i.e. absence of positive feedback) interchangable.

We are concerned with how consistent users are at providing relevance feedback. Namely, we want to quantify the degree to which relevance feedback at iteration t is concordant with feedback given at iteration $t-1$. To do this we need to *a*) identify feature weights based on P^t and N^t, *b*) use these weights to score documents in P^{t-1} and *c*) normalise this score based on the documents in D^{t-1}. This normalised score will be simple to interpret: a score of 0 means that there exists no set of documents that score better from iteration $t-1$ using a definition of relevance from iteration t than P^{t-1}. The feedback at iterations t and $t-1$ is, therefore, perfectly consistent. A score of 0.5, however, means that half of the possible sets of documents from iteration $t-1$ appear more relevant based on the current iterations relevance feedback than what was actually selected.

3.2 Feature weighting

Relevance feedback tends to be coarse-grained: we only know which documents were relevant, not which features contributed to that relevance. Based on the content of documents in P^t and N^t, we want to infer the correlation between the presence and absence of each feature, w_i, with relevance feedback. We calculated the phi coefficient, ϕ, for each feature to determine what was correlated with positive and negative documents. For the following 2×2 contingency table:

	$d \in P$	$d \in N$	total
$w_i \in d$	n_{11}	n_{10}	$n_{1\bullet}$
$w_i \notin d$	n_{01}	n_{00}	$n_{0\bullet}$
total	$n_{\bullet 1}$	$n_{\bullet 0}$	n

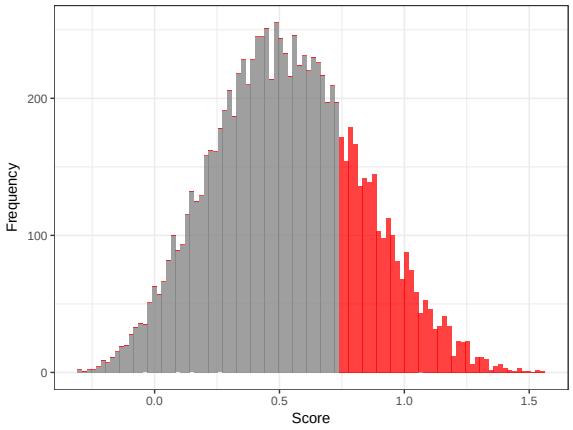

Figure 1: Histogram showing an example permutation score: 24.5% of possible sets of documents from the previous iteration (coloured red) appear more relevant based on current feedback than what was actually selected.

where each word, w_i may or may not be found in a document, d (rows), and d is in set P if it received positive feedback or N if it received negative feedback (columns). Using the counts from the contingency table, we can calculate ϕ_i:

$$\phi_i = \frac{n_{00}n_{11} - n_{10}n_{01}}{\sqrt{n_{1\bullet}n_{0\bullet}n_{\bullet 0}n_{\bullet 1}}} \tag{1}$$

ϕ is related to χ^2 ($\phi^2 = \chi^2/n$), which is commonly used for feature extraction, however, here we want features correlated with documents in P to have a positive weight and, conversely, those in N to have a negative weight. Calculating ϕ_i for all features yields weight vector $\boldsymbol{\Phi}$.

3.3 Relevance feedback inconsistency score

Given the weight vector, $\boldsymbol{\Phi}^t$, derived from releveance feedback at iteration t, we use it to score the set of documents, P^{t-1}, that received positive feedback in iteration $t-1$.

$$score(P^{t-1}, \boldsymbol{\Phi}^t) = \sum_{d_i \in P^{t-1}} \sum_{w_i \in d_i} (w_i \cdot \boldsymbol{\Phi}_i^t), \tag{2}$$

where d_i is $L2$-normalised. In doing so, we are scoring documents from the previous iteration that were deemed relevant, but with a definition of relevance derived from the current iteration. We normalise this score using all permutations of $|P^{t-1}|$ documents from iteration $t-1$ and calculate the proportion of positive document sets that score higher than $score(P^{t-1}, \boldsymbol{\Phi}^t)$:

$$perm(D^{t-1}, P^{t-1}, \boldsymbol{\Phi}^t) = \frac{\sum_{P' \in X} \mathbb{1}(score(P', \boldsymbol{\Phi}^t) > \tau)}{|X|}, \tag{3}$$

where $X = \binom{D^{t-1}}{|P^{t-1}|}$, $\tau = score(P^{t-1}, \boldsymbol{\Phi}^t)$ and $\mathbb{1}(\cdot)$ is an indicator function. If, however, $|X| > 10,000$, then we randomly sample 10,000 positive document sets without replacement from X instead.

Figure 1 demonstrates how the permutation score works. In this example, the user gave positive feedback to 5 documents in iteration

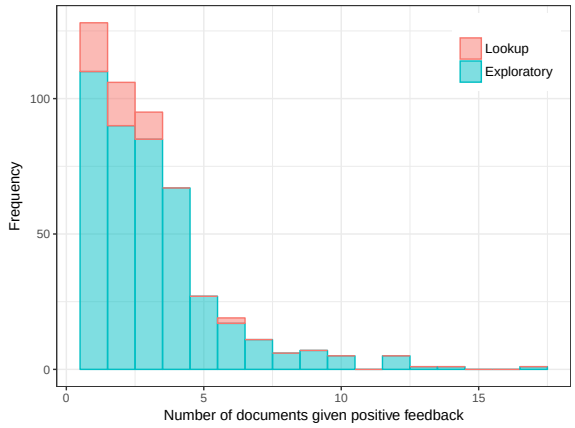

Figure 2: Number of documents given positive feedback in iteration $t-1$ for all consecutive pairs of search iterations used in analysis, stratified by search type.

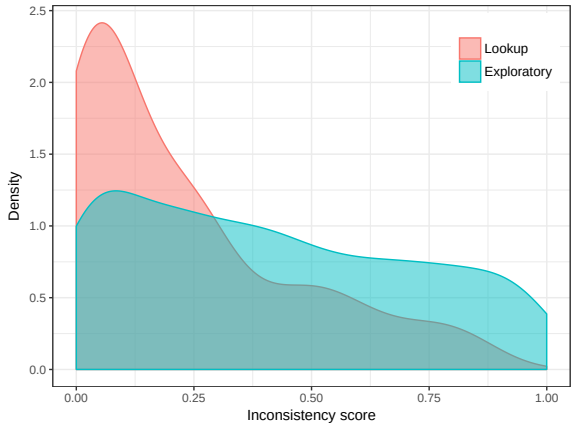

Figure 3: Density plots of inconsistency scores comparing lookup and exploratory searches. The modes for each search task are similar, but feedback during exploratory search is more inconsistent in general.

$t-1$. Based on the weights in Φ^t, these documents scored 0.739. As there are 15,504 combinations of 5 documents from the 20 presented to the user, 10,000 were randomly selected without replacement and scored. The histogram is coloured grey if the score was lower and red if the score was greater than 0.739. As 24.5% scored higher, and would therefore be more consistent with the feedback from the current iteration, the permutation score is 0.245.

4 RESULTS

4.1 Data

We collected data from studies that used the same interactive information retrieval system to perform lookup and exploratory searches [10]. The data set was composed of 188 experiments that were performed by 62 participants. Of the 188 experiments, 30 were lookup searches and 158 were exploratory searches. All the users performing these searches were MSc thesis writers or first year PhD students in computer science or related fields. In the exploratory searches, users were asked to find papers that might be useful for writing their dissertations or other scientific reports, while in lookup search they were asked to find a specific paper they browsed through the day before (without knowing the title or author(s) of the paper). All the searches were performed over the entire arXiv data set of ~1M documents (https://arxiv.org/). We extracted consecutive pairs of iterations where participants gave positive feedback to at least 1 document in both iterations. We found 479 observations (46 lookup, 433 exploratory) that fulfilled this criteria. Figure 2 shows the number of documents that received positive feedback in our 479 observations, stratified by search type. In all experiments, users were presented with 20 documents per iteration. In a majority of cases, users gave positive feedback to 1-4 documents.

4.2 Relevance feedback is more inconsistent in exploratory search than lookup

While relevance feedback from exploratory searches is more inconsistent than lookup search in general, the mode inconsistency for both search tasks is highly similar. Figure 3 shows the distribution of inconsistency scores for lookup and exploratory search tasks. The inconsistency score varies widely; both search tasks had observations at the extremes. Both distributions, however, are skewed to the left, showing that users were more likely to give consistent relevance feedback irrespective of search task. Indeed, the mode inconsistency score is low for both lookup and exploratory search (0.06 and 0.09, respectively).

Users engaged in lookup search give consistent feedback more frequently than those performing exploratory search. If we define a score of 0.1 or lower as *highly consistent*, then relevance judgments from lookup searches are highly consistent 47.8% of the time compared with only 21.5% for exploratory search. These results are consistent with expectations: users performing a lookup search should know what they are looking for and their feedback will, therefore, be more reliable. For exploratory search users, a substantial proportion of relevance feedback is highly inconsistent. For example, for exploratory search, 37.0% of relevance feedback had an inconsistency score of 0.5 or greater, meaning that half of the possible documents selections from previous iteration were more relevant than those actually selected. This is in contrast to lookup search for which only 15.2% of observations scored 0.5 or greater.

4.3 Lookup and exploratory search have different inconsistency distributions

While Figure 3 shows a clear difference between lookup and exploratory search, there is an order of magnitude difference in the number of observations for each search task. According to a Wilcoxon rank sum test, the inconsistency scores generated by lookup and exploratory search tasks come from different distributions ($W = 13639$, $p = 3.742 \times 10^{-5}$). Figure 4 shows the empirical cumulative distribution function (CDF) for lookup search and exploratory search together with the 95% confidence interval. The median inconsistency score for exploratory search was 0.37, whereas for lookup

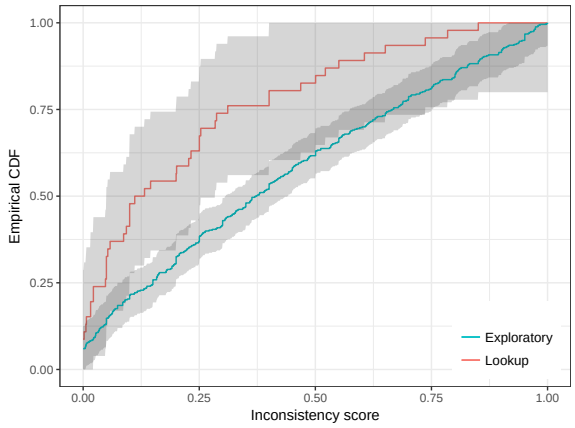

Figure 4: Empirical cumulative distribution function of inconsistency scores comparing lookup and exploratory searches. Shaded bands show the 95% confidence interval.

search it was 0.11. The confidence intervals diverge between 0.1–0.35, which is where the bulk of the difference was found in the density plots (Figure 3), suggesting that the broad trends identified earlier are trustworthy, despite the differences in sample size. While the confidence intervals overlap at (0,0) and (1,1), both CDFs need to pass through those points, so it is unavoidable. We note that multiple observations can come from the same individuals and even search sessions, and are therefore not i.i.d., as is assumed by the Dvoretzky-Kiefer-Wolfowitz inequality used to create the confidence interval for an empirical CDF.

4.4 Inconsistency does not correlate with other experimental variables

Relevance feedback inconsistency is not explained by any of the other variables available in this data set. For both lookup and exploratory search, the inconsistency score is not correlated with the number of documents given positive feedback nor with document diversity (measured using mean pairwise cosine distance between documents). For a limited subset of exploratory search experiments, we collected self-reported knowledge related to the search task on a 1–5 scale from "no knowledge" to "very familiar". Users performed exploratory searches with queries with knowledge level 2, 3 or 4. There was no correlation between knowledge level and the distribution of inconsistency scores, i.e. the inconsistency score distribution for knowledge level 4 was not more similar to those performing lookup search than knowledge level 2 (data not shown).

5 DISCUSSION

Previous work has already demonstrated that user behaviour differs markedly between lookup and exploratory search [2, 3]. In this paper, we demonstrated that relevance feedback is more likely to be inconsistent, as defined by our metric, during exploratory search than lookup search. Our findings suggest that when relevance feedback is used in exploratory search, it should be assumed to contain highly variable measurement error. The inconsistency score we proposed could be used by an IR system to dynamically identify the quality of relevance judgments and weight them in proportion to their consistency in the retrieval algorithm.

In our retrospective analysis we could not properly address why feedback was inconsistent. One hypothesis is that users performing exploratory search are more prone to the diagnosticity effect, i.e. that the surrounding search results have an effect on document relevance [14]. While we did not find any correlation between document diversity and feedback inconsistency (Section 4.4), cosine distance is a poor proxy for users' perception of diversity and to understand whether the diagnosticity effect is a factor would require additional experimentation.

In future work, we want to better understand how users perceive their search goals and, specifically, how they translate those goals into relevance judgments. We need to better understand what users consider to be salient features when they give relevance feedback and whether this can be used to improve user experience. This could be done by either changing how relevance feedback is interpreted by the retrieval algorithm or by helping users to provide the kind of feedback IR systems expect. For example, we could highlight features by their phi coefficient to show users how document-level feedback will be interpreted by the system.

REFERENCES

[1] K. Ahukorala, A. Medlar, K. Ilves, and D. Glowacka. 2015. Balancing exploration and exploitation: Empirical parameterization of exploratory search systems. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*. 1703–1706.
[2] K. Athukorala, D. Głowacka, G. Jacucci, A. Oulasvirta, and J. 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.
[3] K. Athukorala, A. Medlar, A. Oulasvirta, G. Jacucci, and D. Glowacka. 2016. Beyond Relevance: Adapting Exploration/Exploitation in Information Retrieval. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*.
[4] K. Athukorala, A. Oulasvirta, D. Głowacka, J. Vreeken, and G. Jacucci. 2014. Narrow or Broad?: Estimating Subjective Specificity in Exploratory Search. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*.
[5] M. J. Bates. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online review* 13, 5 (1989), 407–424.
[6] A. Chuklin, I. Markov, and M. de Rijke. 2015. Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 7 (2015).
[7] D. Glowacka, T. Ruotsalo, K. Konuyshkova, S Kaski, and G. Jacucci. 2013. Directing exploratory search: Reinforcement learning from user interactions with keywords. In *Proceedings of the 2013 international conference on Intelligent user interfaces*.
[8] J. Y. Kim, M. Cramer, J. Teevan, and D. Lagun. 2013. Understanding how people interact with web search results that change in real-time using implicit feedback. In *Proceedings of the Conference on Information & Knowledge Management*.
[9] G. Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.
[10] 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*.
[11] A. Medlar, J. Pyykkö, and D. Glowacka. 2017. Towards Fine-Grained Adaptation of Exploration/Exploitation in Information Retrieval. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*.
[12] P. Pirolli and S. Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.
[13] I. Ruthven and M. Lalmas. 2003. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review* 18 (2003), 95–145.
[14] A. Tversky. 1977. Features of similarity. *Psychological review* 84, 4 (1977), 327.
[15] R. W. White and R. A. Roth. 2009. Exploratory search: Beyond the query-response paradigm. *Synthesis lectures on information concepts, retrieval, and services* (2009).
[16] X. S. Zhou and T. S. Huang. 2003. Relevance feedback in image retrieval: A comprehensive review. *Multimedia systems* 8, 6 (2003), 536–544.