Evaluation of Rich and Explicit Feedback for Exploratory Search

Esben Sørig
Goldsmiths, University of London
esben.sorig@gold.ac.uk

Nicolas Collignon
University of Edinburgh
nccollignon@gmail.com

Rebecca Fiebrink
Goldsmiths, University of London

Noriko Kando
National Institute of Informatics, Japan

ABSTRACT

A user's goals and interests during an exploratory search task are often ambiguous and complex. When engaging with new documents, people regularly use annotations to help better understand them and make them their own. These annotations can also provide rich information to gain insight into a reader's interests. In particular, it is possible to use highlights as a richer form of feedback compared to traditional document-level relevance feedback. We first show that this form of feedback leads to improvements in document retrieval in exploratory search tasks with simulated users when compared to relevance feedback. We then present an evaluation platform which will allow us to understand the retrieval performance, user experience, and behavioral characteristics of human subjects using highlights as feedback. Finally, we propose an experimental design with human subjects. We hope that our experimental findings will help improve current simulated user evaluations for such systems.

KEYWORDS

exploration, relevance feedback, interactive search, HCI, interactive machine learning, evaluation methods

ACM Reference Format:
Esben Sørig, Nicolas Collignon, Rebecca Fiebrink, and Noriko Kando. 2019. Evaluation of Rich and Explicit Feedback for Exploratory Search. In *WEPIR '19: Workshop on the Evaluation of Personalisation in Information Retrieval, March 14, 2019, Glasgow, UK*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Exploratory search is an essential part of every individual's journey of lifelong learning, whether this is when planning a holiday in a foreign country or starting a new academic research project. Here, we use the term *exploratory search* to describe any type of information search that is open-ended, persistent, and multi-faceted, where the initial goal of the user is complex and initially vaguely defined [12]. The information needs of learners delving into domains of knowledge over extended periods of time, people confronted with chronic illness, work teams designing complex solutions, families making long-term plans, or scientists investigating complex phenomena are poorly supported by existing Web search engines [13]. The role of Exploratory Search Systems (ESS) is to address this shortcoming by providing guidance in exploring complex and unfamiliar information landscapes and to empower users to go beyond single-session look-up tasks [19].

The difficulty for users to express and even know their information needs is a key problem in complex search. However, users usually find it easy to judge the relevance of retrieved documents [2]. Relevance feedback is a well-known information retrieval approach that takes advantage of this observation by letting users judge the relevance of retrieved documents, thereby allowing them to refine their information needs and search results. Traditional relevance feedback methods ask the user to explicitly indicate the relevance of returned documents using binary or multi-scale judgments. Relevance feedback has been shown to improve search performance [15].

Explicit feedback requires users to spend effort engaging in activities that are not part of their usual search behavior, and the benefits are not always apparent. In general, research has pointed towards low user engagement with relevance feedback during general search engine use [18]. Finding the right balance between explicit control and cognitive load is clearly not easy [20]. Recent research in information retrieval has thus focused on inferring relevance from user behavior, since implicit feedback can easily be collected in large quantities and can be gathered at no extra cost to the user [9]. However, many of the studies motivating implicit relevance feedback have not focused on exploratory search, a setting in which the strenuous aspects of providing explicit feedback might be compensated by their benefits. In fact, many researchers still prefer to print hard copies of source materials in order to read and annotate them. Golovchinsky et al. argued that this highlights some of the limits that current ESSs have because they maintain potentially disruptive barriers between the different stages of information retrieval: Searching for documents, printing and reading, and iterating on the search [5].

Interacting with a document can help enhance the reader's understanding, or recall, of the information (a practice sometimes referred to as *active reading* [1]). Annotating documents is therefore a natural part of the exploratory search process for many users. This usually entails highlighting key passages of text and making notes in the margins of the document[4]. Today most of this rich annotation information is lost, either on the physical copy of a document or simply because digital systems do not allow for such annotations. In their study, Golovchinsky et al. showed that feedback based on passages highlighted by users produced significantly better results than relevance feedback for precision and

WEPIR '19, March 14, 2019, Glasgow, UK

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

recall [5]. However, the study was not done in an interactive setting. Participants were shown the same manually selected relevant and irrelevant documents for each task, and they did not see the updated results based on their feedback. This space for further research has, to our knowledge, mostly been overlooked. Interactive query expansion uses relevance feedback to predict a list of relevant terms from which the user selects terms to refine their results [10]. The technique allows users to control the effect of their relevance feedback but is limited to keyword selection from an automatically generated list. Recent research in dynamic information retrieval has shown promising results with algorithms using annotation-level feedback when used by simulated users [21]. The necessity of evaluating these algorithms with more human-like simulations has been emphasized [6]. Indeed, the algorithm evaluations have only been done on a sequence of iterations from a single query [21] and remain far from *real world* exploratory search. To our knowledge, no study has evaluated the effects of annotation feedback on performance and user experience during exploratory search tasks with real participants, in a multi-query setting.

We have developed an evaluation platform to test the effect of annotation feedback on search performance, user experience and user behaviour. Our platform allows users to search and iteratively refine their results based on the text they highlight in returned documents. In section 2, we firstly present our approach to personalizing search results based on user annotations. In section 3, we present encouraging performance results of our approach compared to relevance feedback using user simulation. To test whether annotation feedback leads to similar performance improvements when used by real users, we have developed an experimental design which we present in section 4. In section 5, we present the methods for our experimental data collection. In section 6, we present our evaluation methods for both search performance and user experience.

2 USING ANNOTATIONS FOR SEARCH

The search system in our evaluation platform ranks documents based on user queries and either user annotations (experimental condition) or document-level relevance feedback (control condition). A multinomial Naive Bayes classifier is used to model relevant and irrelevant documents. Documents are ranked by the probability of being relevant under the model. The model is trained on user queries, annotation feedback, and relevance feedback. We use document-level relevance feedback using the ordinary supervised learning approach. Rather than labels (i.e. *relevant* or *not relevant*) on entire documents, queries and annotation feedback are feedback on specific features of a document (i.e. words are labeled rather than the document). We use this type of feedback in a similar way to [17] and update the prior distribution for the features selected by the query or annotation in the relevant class. We pre-train both relevant and irrelevant classes on the entire corpus to avoid overfitting to the limited amount of feedback the user provides.

In cases where the corpus size makes a real-time prediction with the multinomial Naive Bayes classifier impossible, the system computes an initial ranking using the Rocchio algorithm [15]. The top N documents from the Rocchio ranking are then classified using the multinomial Naive Bayes model (in our simulation experiment we let $N = \infty$ while we use $N = 1000$ in our experiment with human subjects).

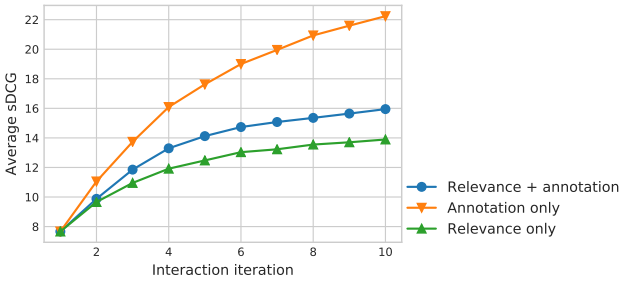

Figure 1: Average session discounted cumulative gain across all 60 search tasks computed each time the search results are updated based on feedback from the previous iteration's search results (interaction iteration).

3 SIMULATION EXPERIMENT

We have evaluated the performance of our system with each feedback type using the TREC 2017 Dynamic Domain Track user simulator [21] on the New York Times (NYT) Annotated Corpus [16]. The corpus contains over 1.8 million articles published in NYT between 1987 and 2007. The user simulator performs 60 different search tasks, each consisting of an initial query and multiple iterations of feedback and updates to the search results. On each iteration, our system returns 5 documents to the simulator. Based on ground truth annotations of passages relevant to each task, the simulator returns all relevant passages contained in the current search results including a rating of how relevant each passage is (on a scale from 1 to 3). Furthermore, the simulator provides binary document-level relevance feedback which marks all documents with relevant passages as relevant and all other documents as irrelevant. We let the simulator perform 10 iterations of feedback (i.e. the simulator gives feedback on 50 documents in total).

We ran the simulation under three conditions: only using annotation feedback, only using relevance feedback, and using both in combination. The ranking performance was measured using Expected Utility, Cube Test, and Session Discounted Cumulative Gain (sDCG) averaged over all search tasks [21]. On all metrics, annotation feedback outperforms relevance feedback and the combination. The sDCG performance is shown in Figure 1. We hypothesize that the improvement in performance when using annotation feedback comes from the fact that annotations capture the key aspects that make a document relevant, whereas relevance feedback captures the statistical nature of relevant documents. Relevance feedback therefore is a noisier signal. Using relevance feedback or combining annotation feedback and relevance feedback, the simulated user is required to give feedback on more documents to achieve comparable performance to annotation feedback.

Given the improved retrieval performance results observed with user simulation, we want to test if this is also true for real users. To study the difference in search performance, user experience, and user behavior when providing annotation-level vs document-level feedback, we will evaluate the system using an experiment with human subjects. We believe a more human-centered approach to

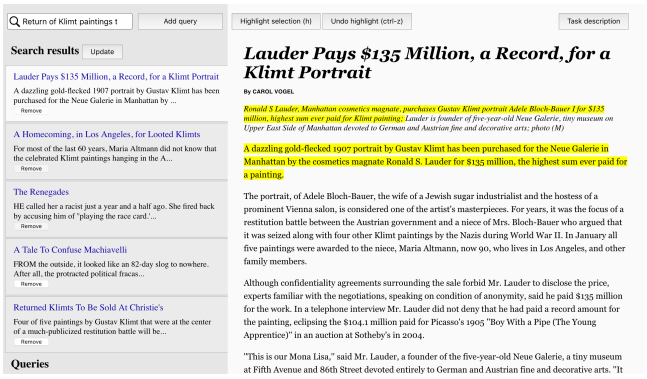

Figure 2: User interface for the annotation condition. The query box is located in the top left corner. Search results are displayed on the left. The currently selected article is displayed on the right along with buttons for highlighting the document. Below the results are lists of queries entered so far ("Queries"), articles that have previously been annotated ("Annotated Articles"), and articles that have been removed from search results ("Removed Articles").

search system evaluations is important to better understand how people interact with an annotation-level feedback mechanism, and hope the experimental results can help design more human-like, simulated-user evaluation tools. In the following parts, we present our experimental design, methods and evaluation methods.

4 EXPERIMENT WITH HUMAN SUBJECTS

4.1 The tasks

The experiment will consist of a sequence of three search tasks. During each search task, participants will have to find relevant information by searching for documents, read and highlight relevant passages, and judge documents related to the given topic.

We select tasks from the TREC-17 Dynamic Domain track. Topics determine the domain of the search task. Each of the 60 TREC-17 search topics consists of a paragraph that characterizes relevant documents. Each topic contains a number of relevant passages that were discovered and graded by expert human annotators according to relevance to the topic. All relevance judgments are at a passage level, and each document may have several passages that relate to a topic. The relevance scores allow us to compute different evaluation metrics.

For the experiment, we choose 6 topics of different complexity to understand how the search system will influence user experience and search performance for different types of exploratory search tasks. Finding measures to estimate task complexity has been a longstanding research question [7] and perceived user difficulty is still the best indicator for the complexity of tasks [8]. In this work, we use the results of the user simulation as an objective measure for expected task complexity, from the TREC Dynamic Domain track [21].

4.2 Conditions

Document-level condition. In the first condition, participants will provide document-level feedback to the search system by marking articles as relevant or irrelevant. We refer to this condition as the *DL*-condition.

Annotation condition. In the second condition, participants will provide feedback at the paragraph, sentence, or word level, by highlighting relevant text. We refer to this condition as the *A*-condition.

In both conditions, participants will engage in the search process much like one would in a typical web search. Participants are free to give as much or little feedback (relevance and annotation respectively) as they want. However, participants will be encouraged to make use of the respective feedback features. They are free to add, edit, or remove previous feedback, add or remove queries to/from the search, and update their search results at any point.

4.3 The Search Interface

We designed the search interface so participants in both the *DL*-condition and the *A*-condition would have an experience as close to the current web exploratory search experience as possible (see Figure 2). During the search process, users are presented with 5 documents at a time. Participants can update the list as often as they want after providing feedback to the system. Once a document has been marked as relevant, marked as irrelevant (DL-condition), or highlighted (A-condition), it does not show up in the search results again but appears in the "Annotated Articles" list instead. As we do not require participants to provide feedback on every document, they may move documents to the "Removed Articles" list to avoid them from appearing in future results. Users may edit previously provided feedback and queries at any time. Providing such explicit control over feedback to adapting systems has been shown to improve performance and user experience in other domains [11]. Participants are told before the task that highlights (A-condition) and relevance marking of documents (DL-condition) will help yield better document sources. Participants are asked to type in an initial query after they have read the task description.

5 METHODS

Participants

We will recruit 80 participants via Amazon Mechanical Turk who will receive $8 plus a performance-dependent bonus of up to $5 as a reward based on their search performance. Each MTurk Human Intelligence Task (HIT) will require participants to complete a full set of three search tasks, and they will only be allowed to complete one HIT. To help ensure English language proficiency and quality control, we will only recruit MTurk workers located in the U.S. with a $\geq$ 95% acceptance rate.

Design

In the experiment, participants will be told they are journalists applying for a job at the New York Times. In the interview, they have to gather information from the NYT archive for different tasks. They are told that the documents they find will be compared with expert ratings and that their performance will be tied to a reward bonus. In each condition, participants are told how their respective feedback types (*document-level* and *annotation-level*) influences the ranking of documents throughout the search session. We employ

a between-subjects design and randomly allocate each participant to one of the conditions. The three tasks are randomly allocated to each participant.

Behavioral data collection

During the experiment, we will collect all feedback actions (highlights, relevance feedback, and removed articles) with time stamps. We will also collect overall time on spent on each task, the time spent on each article and dwell time on the result list (estimated with mouse movement and scrolling), as well as participant answers to the pre- and post-task questionnaires.

6 EVALUATION METHODS

6.1 Retrieval performance

We will measure the retrieval performance of participants under the two experimental conditions using the same multi-session ranked retrieval metrics used in the user simulation.

6.2 User Experience

In addition to the retrieval performance of participants, we will compare the two conditions by contrasting answers of Pre- and Post Task Questionnaires.

6.2.1 Questionnaires. Participants will complete a pre-task questionnaire before starting each task (denoted as *PreTask*) and two questionnaires after completing each task: a post-task counterpart to the pre-task questionnaire (denoted as *PostTask*) and a short form of O'Brien's User Engagement Scale [14] (denoted as (UE-SF). In the questionnaires, participants will be asked to report their level of agreement on a 7-point scale from 1 (strongly disagree) to 7 (strongly agree) with the different statements.

6.2.2 PreTask and PostTask Questionnaires. The PreTask and PostTask questionnaires will consist of the same statements, with the aim of evaluating how participants changed their mind after having gone through the task. The questions in the PreTask and PostTask questionnaires can be categorized according to the themes: (1) prior knowledge/knowledge increase, (2) interest/interest increase, (3) expected/experienced difficulty. These three aspects (knowledge, interest, difficulty) are common in IIR studies [7]. We will include questions about (4) determinability and (5) subjectivity to better understand how specific task features might influence behavior and performance following [3]. We will also include free-form answer boxes to gather qualitative comments participants might have about the task.

6.2.3 User Engagement Scale (UES-SF). We follow the UES-Short Form (UES-SF) by O'Brien et al. [14], and use 7 questions designed to capture four dimensions of engagement: (1) focused attention (FA), (2) perceived usability (PU), (3) aesthetic appeal (AE) and (4) reward (RW). We use a subset of the original questionnaire to limit its effect on user experience and cognitive load.

7 CONCLUSION

Early work on using user annotations for personalization [5] offered supporting evidence in favor of using rich feedback for better exploratory search performance. In this paper, we showed a similar improvement in search performance using the TREC Dynamic Domain user simulator. However, we do not yet understand the effect of annotation feedback on search with real users. We therefore presented an experimental framework which allows us to study the retrieval performance, user experience, and behavioral characteristics of users using such a system. We believe this framework will help us understand whether annotation feedback is beneficial for real users, but also help us design more realistic user simulations to inform the design of better search algorithms in the future.

ACKNOWLEDGMENTS

This work was supported by Microsoft Research through its PhD Scholarship Programme and by the National Institute of Informatics (Japan) through its International Internship Programme and travel support. JSPS KAKENHI Grant Numbers JP16H01756.

REFERENCES

[1] Mortimer J Adler and Charles Van Doren. 1972. How to Read a Book, rev. ed. *New York* (1972).
[2] D. C. Blair and M. E. Maron. 1985. An Evaluation of Retrieval Effectiveness for a Full-text Document-retrieval System. *Commun. ACM* 28, 3 (1985), 289–299.
[3] Robert Capra, Jaime Arguello, Heather O'Brien, Yuan Li, and Bogeum Choi. 2018. The Effects of Manipulating Task Determinability on Search Behaviors and Outcomes. In *Proceedings of ACM SIGIR 18*. ACM, 445–454.
[4] Sally Jo Cunningham and Chris Knowles. 2005. Annotations in an academic digital library: the case of conference note-taking and annotation. In *International Conference on Asian Digital Libraries*. Springer, 62–71.
[5] G. Golovchinsky, M. N. Price, and B. N. Schilit. 1999. From reading to retrieval: freeform ink annotations as queries. In *Proceedings of ACM SIGIR 99*. ACM, 19–25.
[6] Evangelos Kanoulas, Leif Azzopardi, and Grace Hui Yang. 2018. Overview of the CLEF Dynamic Search Evaluation Lab 2018. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 362–371.
[7] Diane Kelly et al. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval* 3, 1–2 (2009), 1–224.
[8] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. 2015. Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework. In *Proceedings of ICTIR '15*. ACM, 101–110.
[9] Diane Kelly and Jaime Teevan. 2003. Implicit feedback for inferring user preference: a bibliography. In *Acm Sigir Forum*, Vol. 37. ACM, 18–28.
[10] Jürgen Koenemann and Nicholas J. Belkin. 1996. A Case for Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness. In *Proceedings of CHI '96*. 205–212.
[11] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of IUI '15*. 126–137. https://doi.org/10.1145/2678025.2701399
[12] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.
[13] Gary Marchionini and Ryen W White. 2009. Information-seeking support systems [guest editors' introduction]. *Computer* 42, 3 (2009), 30–32.
[14] H. L. O'Brien, P. Cairns, and M. Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (2018), 28–39.
[15] I. Ruthven and M. Lalmas. 2003. A Survey on the Use of Relevance Feedback for Information Access Systems. *Knowl. Eng. Rev.* 18, 2 (June 2003), 95–145.
[16] Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia* 6, 12 (2008), e26752.
[17] Burr Settles. 2011. Closing the Loop: Fast, Interactive Semi-supervised Annotation with Queries on Features and Instances. In *Proceedings of EMNLP '11*. 1467–1478.
[18] A. Spink, B. J. Jansen, and H. Cenk Ozmultu. 2000. Use of query reformulation and relevance feedback by Excite users. *Internet research* 10, 4 (2000), 317–328.
[19] Ryen W White and Resa A Roth. 2009. Exploratory search: Beyond the query-response paradigm. *Synthesis lectures on information concepts, retrieval, and services* 1, 1 (2009), 1–98.
[20] Max L Wilson, MC Schraefel, and Ryen W White. 2009. Evaluating advanced search interfaces using established information-seeking models. *Journal of the American Society for Information Science and Technology* 60, 7 (2009), 1407–1422.
[21] Grace Hui Yang, Zhiwen Tang, and Ian Soboroff. 2017. TREC 2017 Dynamic Domain Track Overview.. In *TREC*.