Predicting Retrieval Success Based on Information Use for Writing Tasks

Pertti Vakkari[1][0000−0002−4441−5393], Michael Völske[2][0000−0002−9283−6846], Martin Potthast[3][0000−0003−2451−0665], Matthias Hagen[4], and Benno Stein[2]

[1] University of Tampere, FIN-33014, Finland
pertti.vakkari@uta.fi
[2] Bauhaus-Universität Weimar, 99423 Weimar, Germany
<first>.<last>@uni-weimar.de
[3] Leipzig University, 04109 Leipzig, Germany
martin.potthast@uni-leipzig.de
[4] Martin-Luther-Universität Halle-Wittenberg, 06108 Halle, Germany
matthias.hagen@informatik.uni-halle.de

Abstract. This paper asks to what extent querying, clicking, and text editing behavior can predict the usefulness of the search results retrieved during essay writing. To render the usefulness of a search result directly observable for the first time in this context, we cast the writing task as "essay writing with text reuse," where text reuse serves as usefulness indicator. Based on 150 essays written by 12 writers using a search engine to find sources for reuse, while their querying, clicking, reuse, and text editing activities were recorded, we build linear regression models for the two indicators (1) number of words reused from clicked search results, and (2) number of times text is pasted, covering 69% (90%) of the variation. The three best predictors from both models cover 91-95% of the explained variation. By demonstrating that straightforward models can predict retrieval success, our study constitutes a first step towards incorporating usefulness signals in retrieval personalization for general writing tasks.

1 Introduction

In assessing information retrieval effectiveness, the value of search results to users has gained popularity as a metric of retrieval success. Supplementing established effectiveness indicators like topical relevance [1–3], the worth [4], utility [2], or usefulness [1] of information depends on the degree to which it contributes to accomplishing a larger task that triggered the use of the search system [4, 2, 5]. Despite the growing interest in information usefulness as a retrieval success indicator, only a handful of studies have emerged so far, and they typically focus on perceived usefulness rather than on the actual usage of information from search results. Even fewer studies explore the associations between user behavior and information usage during task-based search. The lack of contributions towards this important problem arises from the difficulty of measuring the usefulness of a search result with regard to a given task in a laboratory setting.

The final authenticated version is available online at https://doi.org/10.1007/978-3-030-00066-0_14

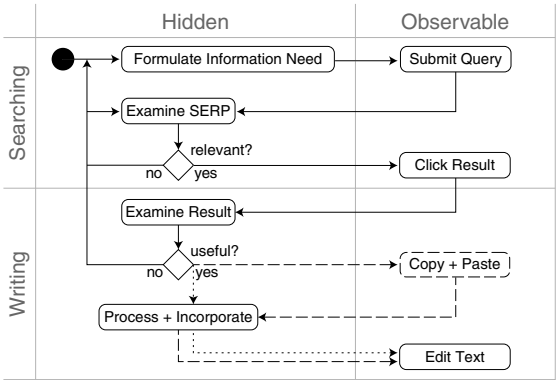

Fig. 1. User actions in the search and writing process: our study design involving text reuse (dashed lines) allows for more direct observation of users' usefulness assessments of search results than would be possible without reuse (dotted lines).

Focusing on essay writing, with this paper, we lift essay writing for the first time into the realm of fine-grained usefulness quantification. Using the task "essay writing with text reuse" as a surrogate, the usefulness of a search result becomes directly observable as a function of copy and paste (see Figure 1). Furthermore, by analyzing a large corpus of essays with text reuse, where the search and writing behavior of their writers has been recorded, we identify two specific usefulness indicators based on text reuse behavior and build linear regression models that predict result usefulness based them. Keeping the limitations of our approach in mind, we believe that these results offer promising new directions for the development of search systems that support writing tasks at large.

In what follows, Section 2 reviews related work, Section 3 outlines our methodology, and Section 4 reports on our models. In Section 5 we discuss the consequences and potential limitations of our work.

2 Related Work

Search results are useful if information they contain contributes to the task that triggered information searching [5]. Users are expected to click, scan, and read documents to identify useful pieces of information for immediate or later use [6]. Only a few studies on the usefulness of search results focus on predicting the usefulness for some task [7–10], while most others are more interested in comparing relevance and usefulness assessments (e.g. [11]).

In most cases, usefulness is operationalized as perceived by users, not as the actual usage of information. Kelly and Belkin [7] explored the association between documents' display time and their usefulness for the retrieval task. Here, usefulness was operationalized as the degree of users' belief of how helpful the document was in understanding and completing a particular task. They found no association between usefulness and dwell time, regardless of the task type.

Liu and Belkin [8] also studied whether the time spent on a clicked document was associated with its perceived usefulness for writing a journalistic article. They found a positive association between the dwell time on a document and its usefulness assessment. Users typically moved back and forth between the text they produced and the document informing their writing. In the context of the essay writing task of the scenario we explore, the copy-pasting of (parts from) search results may be an even more direct relation. Liu et al. [9] later modeled users' search behavior for predicting the usefulness of documents: they had users assess the usefulness of each saved page for an information gathering task, and employed binary recursive partitioning to identify the most important predictors of usefulness. In an ascending order, dwell time on documents, time to the first click, and the number of visits on a page were the most important predictors—the longer the dwell time, the more visits on a page and the shorter the time to first click, the more useful the page. Mao et al. [10] recently modeled the usefulness of documents for answering short questions by content, context, and behavioral factors, where usefulness was measured on a four-point scale. They found that behavioral factors were the most important in determining usefulness judgments, followed by content and context factors: the perceived usefulness of documents was positively correlated with dwell time and similarity to the answer, and negatively with the number of previous documents visited.

By comparison, Ahn [12] and He [13] evaluated the actual usefulness of information retrieved by measuring to what extent search systems support finding, collecting and organizing text extracts to help answer questions in intelligence tasks, with experts assessing the utility of each extract. Sakai and Dou [14] proposed a retrieval evaluation measure based on the amount of text read by the user while examining search results, presuming this text is used for some purpose during the search session.

3 Experimental Design

We base our investigation on a dataset of 150 essays and associated search engine interactions collected by Potthast et al. [15] as part of a large-scale crowdsourcing study with 12 different writers (made available to the research community as the Webis Text Reuse Corpus 2012,[5] Webis-TRC-12). We briefly review key characteristics of the Webis-TRC-12 and its collection process below, before detailing our conceptualization of document usefulness, and the variables we derive for our study.

3.1 Data

Each of the dataset's essays is written in response to one of 150 writing prompts derived from the topics of the TREC Web tracks 2009–2011. The writers were instructed to use the ChatNoir search engine [16] indexing the ClueWeb09 web

[5] https://webis.de/data/webis-trc-12.html

crawl to gather material on their essay topic; all submitted queries and visited search results were recorded. For the purpose of writing the essay, the corpus authors provided an online rich text editor, which logged the interactions of writers with their essay texts in fine-grained detail, by storing a new revision of the text whenever a writer stopped typing for more than 300ms.

The 12 writers were hired via the crowdsourcing platform Upwork, and were instructed to choose a topic to write about, and produce an essay of 5000 words using the supplied search engine and rich text editor. Writers were encouraged to reuse text from the sources they retrieved, paraphrasing it as needed to fit their essays. As reported by the corpus authors [15], the writers were aged 24 years or older, with a median age of 37. Two thirds of the writers were female, and two thirds were native English speakers. A quarter each of the writers were born in the UK and the Philippines, a sixth each in the US and India, and the remaining ones in Australia and South Africa. Participants had at least two and a median of eight years of professional writing experience.

The fine-grained data collection procedure, along with the intermeshing with established datasets like the TREC topics and ClueWeb09, has enabled the search and writing logs in the Webis-TRC-12 to contribute to several research tasks, including the study of writing behavior when reusing and paraphrasing text [15], of plagiarism and source retrieval [17], and of search behavior in exploratory tasks [18]. In the present study, we examine writers' search and text collection behavior to predict the usefulness of retrieved documents for the underlying essay writing task.

3.2 Operationalizing the Usefulness of Search Results

We limit our conception of usefulness to cover only information usage that directly contributes to the task outcome in form of the essay text, and exclude more difficult to measure "indirect" information usage from our consideration, such as learning better query terms from seen search results.

Usefulness implies that information is obtained from a document to help achieve a task outcome. In the following, we quantify usefulness by focusing only on cases where information is directly extracted from a document, not where it is first assimilated and transformed through the human mind to form an outcome. According to our definition, information is useful if it is extracted from a source and placed into an evolving information object to be modified. In the context of essay writing with text reuse, this means that information is copied from a search result and pasted in the essay to be written.

We measure the usefulness of documents for writing an essay in two dimensions, both based on the idea that a document is useful if information is extracted from it. First, we measure the number of words extracted from a document and pasted into the essay—this measure indicates the amount of text that has the potential to be transformed as a part of the essay. Second, we quantify usefulness as the number of times any text is pasted per clicked document.

The limitations of these measures include that they do not reflect the possible synthesis of pasted information or the importance of the obtained passage of text.

Table 1. Means and standard deviations of study variables (n=130).

Query Variables	μ	σ	**Click Variables**	μ	σ
Queries	46.5	41.9	Clicks per query	4.0	4.6
Unique queries	24.9	18.1	Click trails per query	1.8	1.4
Anchor queries	5.2	6.1	% Useful clicks per query	30.8	12.9
Querying time per query	53.1	46.1	Result reading time per paste	262.0	357.7
% Unique queries of all queries	62.6	20.8	... per click per query	48.9	30.4
% Anchor queries of all queries	10.5	7.1	... per major revision	172.4	141.7
Query terms per query	5.4	1.5	*Text Editing Variables*		
Unique query terms (UQT) per query	0.8	0.4	Writing time per major rev.	867.3	666.0
UQT from documents per query	0.6	0.3	Revisions per paste	175.7	225.7
% UQT from snippets	78.6	8.7	Writing time per paste	1270.4	1100.5
% UQT from docs	67.1	20.8	Words in the essay	4988.1	388.8
% UQT per query	15.9	7.6	*Other Independent Variables*		
UQT per unique query	1.3	0.4	Search sessions	7.4	4.1
Dependent Variables					
Words per useful click per query				325.0	420.8
Pastes per useful click per query				1.2	1.8

It is evident that the amount and importance of information are not linearly related, although users were allowed to use the pasted text directly for the essay without originality requirements. Our idealization excludes the qualitative aspects of information use; the presupposition that an increase in the amount of pasted text reflects usefulness directly resembles typical presuppositions in information retrieval research: for instance, Sakai and Dou [14] suppose that the value of a relevant information unit decays linearly with the amount of text the user has read. In general, a similar supposition holds for the DCG measure. These presuppositions are idealizations that we also apply in our analyses.

3.3 Independent Variables

For predicting the usefulness of search results, we focus on query, click, and text editing variables to build linear regression models. Temporally, querying and clicking precedes the selection of useful information, while the usage and manipulation of information succeeds it. Since we use aggregated data over all user sessions, we treat the search and writing process as a cross-sectional event, although querying, clicking, and text editing occur over several sessions. Since the editing of the essay text is connected with querying and clicking in a session, it is important to take into account also text editing variables in analyzing the usefulness of search results over all sessions. Therefore, we also select aggregated text editing variables for our models, although this solution is not ideal in every respect for representing the temporal order of the process.

Based on previous studies [9, 10, 18], we select 13 query variables, 6 click variables, 4 text editing variables, and 1 other variable, yielding the 24 variables in total depicted in Table 1. Here, anchor queries refer to those queries repeatedly revisited throughout a session, in order to keep track of the main theme of the

Independent Variable	Group	β	R^2 Change
Clicks per Q	C	0.38***	0.400
Revisions per paste	T	−0.15*	0.144
Unique queries	Q	−0.34***	0.081
Search sessions	O	−0.18**	0.037
Words in the essay	T	0.12*	0.019
Result reading time per paste	C	−0.24***	0.010

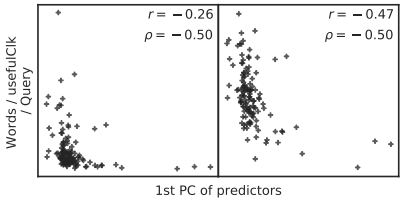

Fig. 2. Left: Regression model for the number of words pasted per useful click per query (n=130), with predictors from the (Q)uery, (C)lick, (T)ext editing and (O)ther variables. Right: First principal component of this model's predictors and dependent variable; effect of the latter's log-transform on Pearson (r) and Spearman (ρ) correlation.

task; time spent querying, reading, and writing is measured in seconds; a click trail begins on the search result page, potentially following further links in the result document. We build regression models for both dependent variables and apply a stepwise entering method of predictors [19].

Regression analysis requires linearity between independent and dependent variables but in our case, the associations of both measures of usefulness with the major independent variables turned out to be non-linear—as evidenced by a large discrepancy between the Pearson and Spearman correlation coefficients, shown in the right-hand plots in Figures 2 and 3. Therefore, we logarithmically transformed both words per useful click per query, and pastes per useful click per query (base of 10), enhancing linearity notably (Figure 2, right). The predictor result reading time per click per query still showed a non-linear association, and was log-transformed as well (Figure 3, right).

While the writers were instructed to produce essays of about 5000 words, some essays were notably shorter or longer [18]. We excluded essays shorter than 4000 and longer than 6000 words from the analysis, as well as four essays with missing variables, yielding 130 observations in total.

4 Results

Based on the variables we derive from the dataset, we investigate two linear regression models of document usefulness—each predicting one of the dependent variables at the bottom of Table 1. The first model uses the number of words pasted per useful click per query as dependent variable, using the amount of text extracted as an indicator of a document's usefulness, whereas the second model quantifies usefulness as the number of times text was extracted, using the number of pastes per useful click per query as dependent variable.

4.1 The Number of Words Pasted per Document

The model is significant (R^2=.703; Adj R^2=.688; F=48.4; p<.000) consisting of six predictors. It explains 68.8% of the variation in the number of words pasted per useful click per query. The tolerance of all variables is greater than .60. The

Independent Variable	Group	β	R^2 Change
Clicks per Q	C	0.33***	0.484
Writing time per paste (sec)	T	−0.42***	0.210
Queries	Q	−0.18**	0.162
Unique queries	Q	−0.24***	0.013
% Useful Clicks per Q	C	0.18***	0.008
Result reading time per C per Q	C	−0.15***	0.015
% Anchor Q of all Q	Q	0.12***	0.006
% Unique Q of all Q	Q	0.14**	0.007
% UQT from snippets of all UQT	Q	0.07*	0.004

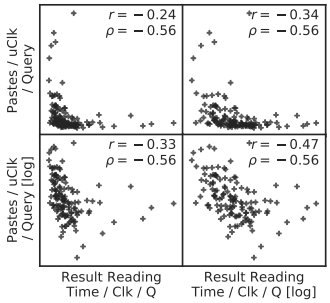

Fig. 3. Left: Regression model for the number of pastes per useful click per query (n=129; groups as in Figure 2, left). Right: Scatter plots of "Result reading time per click per query" against the dependent variable, each before and after log-transform.

four strongest predictors—the number of clicks per query, the number of revisions per paste, the number of unique queries and the number of search sessions—cover 66.2 percentage points of the variation in the number of words pasted (Figure 2, left). The remaining two variables cover 2.9 percentage points of it. Limiting the model to the four major factors, it is possible to reach an accuracy of two thirds in predicting document usefulness.

As per the model coefficients shown in Figure 2, left, the more clicks users make per query, and the less time they spend reading result documents per paste, the more words are pasted per click per query. The number of revisions per paste reduces the number of words pasted. Increases in the number of search sessions and in the number of unique queries reduce the amount of text pasted, while an increase in the number of words in the essay increases the amount of text pasted. The number of unique queries and the number of search sessions are partially correlated, but contribute to the model in this case. Further, fewer clicks per query, more time reading documents, and a greater number of revisions per paste are associated with a smaller amount of text pasted. We hypothesize that difficulties in formulating pertinent queries lead to voluminous querying, and to a greater number of search sessions, which lead to fewer clicks, to longer dwell times per paste, and to a greater number of revisions per paste, all contributing to a smaller number of words pasted.

4.2 The Number of Pastes per Document

The regression analysis produces a model with nine variables significantly predicting the number of pastes per useful click per query. The model is significant (R^2=.908; Adj R^2=.902; F=131.2; p<.000) and covers 90.2% of the variance in the number of pastes per document (Figure 3, left). The three most important predictors—the number of clicks, the writing time per paste, and the number of queries—together explain 85.6% (R^2 Change) of the variation in the number of pastes; the remaining six predictors cover 4.6 percentage points of variation.

The direction of effect in click, query and text editing variables differs: Increasing values of click variables—except reading time—increase the chance that documents provide material for the essay. The query variables both increase and decrease the chance of finding useful documents, while an increase in writing time per paste decreases that chance. Compared to the previous model, click variables have a proportionally smaller contribution compared to query variables, while the relative contribution of text editing variables remains on about the same level. The direction of effect in predictors remains similar; the content of the model essentially resembles the previous one, although some predictors change: writing time per paste resembles revisions per paste, while result reading time per click per query resembles result reading time per paste. The proportions of anchor queries and unique queries are new predictors compared to the previous model.

The model indicates that the more clicks per query, the larger the proportion of useful clicks of all clicks and the shorter the dwell time in clicked documents per query, the more useful the retrieved documents are. Increases in the number of queries and unique queries decrease the usefulness of clicked documents, while increases in the proportion of anchor queries and unique queries of all queries increase the chance that documents are useful.

Multicollinearity tolerance is the amount of variability of an independent variable (0-1) not explained by the other independent variables [19]. Five out of the nine predictors in the model were query variables. Tolerances of the number of queries (.240), the number of unique queries (.291) and the proportion of unique queries (.394) indicate that they depend quite heavily on other variables in the model. Therefore, leaving only the number of queries to represent querying would be reasonable and make the model more parsimonious.

We may conjecture that a smaller number of unique queries with good keywords from snippets produce a good result list. This contributes to a proportionally larger number of useful documents that require less dwell time for obtaining needed information for the essay. The information pasted is pertinent, not requiring much time to edit to match the evolving text. Naturally, the validity of this hypothetical process remains for later studies to test.

4.3 Comparing the Models

The explanatory power of the model predicting the number of words pasted is weaker, covering 68.8% of the variation in document usefulness, while the model for pastes covered about 90% of the variation.

The contributions of query, click and text editing variables vary between the models (Table 2). The relative effect ($\sum R^2$ Change) of click variables is notably greater in both models compared to query and text editing variables. Text editing variables have a somewhat greater role compared to query variables in predicting usefulness as indicated by the number of words pasted. Also the number of search sessions and the number of words in an essay have a minor impact on potential document usefulness. The models have only two predictors in common: the number of unique queries and the number of clicks per query.

Table 2. Summary of models: number of predictors, and relative importance ($\sum R^2$ change), per variable group for both models of search result usefulness.

Characteristics	**Number of Words**	**Number of Pastes**
Adj R^2	0.688	0.902
# Variables ($\sum R^2$ Change)	6	9
Query	1 (0.08)	5 (0.19)
Click	2 (0.41)	3 (0.51)
Text Editing	2 (0.16)	1 (0.21)
Other (sessions)	1 (0.04)	-

In each model, three variables cover over 90% of the explained variation in document usefulness, one of them being a query, one a click and one a text editing variable. The most powerful variable is clicks per query in both models. Thus, one could predict each type of document usefulness by a very simple model.

In both models, the number of queries and the number of unique queries have a negative effect on document usefulness, while all proportional query variables have a positive effect. Clicks per query have a positive contribution to usefulness, while dwell time has a negative contribution. Number of revisions and writing time per paste both have a negative effect on document usefulness. In the model for the number of pasted words, the number of sessions has a negative effect on usefulness, while the number of words in the essay has a positive effect.

Altogether, it seems that clicked result documents are more useful, if: the user issues fewer queries over fewer sessions, makes more clicks per query, but with shorter dwell time on individual documents, makes fewer revisions to the essay per pasted text snippet, and writes a longer essay. Although regression analysis does not indicate associations between independent variables, we conjecture that users who issue fewer queries have better result lists, click more per query, spend less time reading documents, all this producing more useful documents per click per query. This hypothetical process remains to be tested in future work.

5 Discussion and Conclusions

Our study is one of the first attempts to analyze the usefulness of clicked search results based on information usage, instead of measuring perceived usefulness by asking the user (cf. [12, 13]). The results extend our knowledge about factors predicting the actual usefulness of documents—and thus the user's success at finding useful sources—in the context of longer-lasting tasks like essay writing. Usefulness itself, we model by the number of pastes per useful click (indicating whether a documents contain information used in composing the essay), and by the number of words pasted per click (indicating the potential amount of useful information in documents).

Our regression models cover about 90% of the variation in pastes from clicked search results and about 69% of the variation in the number of words pasted per clicked search result. We argue that the number of pastes and the number of

pasted words reflect the actual usefulness of search results fairly validly: for the writers we study, pasting precedes usage in the final essay [18].

We also observe that increased search result usefulness is associated with decreasing effort to edit the pastes for the essay. This is likely a result of the fact that writers were explicitly permitted to reuse text from the sources they found, without having to think about originality requirements. Hence, provided they found appropriate sources, writers could place passages from search results directly as a part of their essays. If the usefulness indicators reflect authors finding sources that require little editing, they should be correlated with less editing of pastes. To test this hypothesis, we measure the proportion of reused words out of all words in the essay (authors annotated the text they reused themselves, as part of the original study); it can be reasonably assumed that the higher this proportion, the less the pasted text is edited. We find that Spearman correlations of the proportion of reused words with the number of pastes (ρ=.27**) and the number of words pasted (ρ=.18*) are significant. Thus, decreasing effort in editing pasted text reflects the usefulness of pastes in composing the essay.

Further, our models indicate that the fewer queries a user makes, the more clicks per query, and the less text editing takes place, the more useful the search results are. This matches well with previous findings: an increase in clicks has been shown to correlate with search satisfaction [20] and the perceived usefulness of documents [9]. However, our results also show that an increase in dwell time decreases search result usefulness. This contradicts many earlier findings that dwell time is positively associated with usefulness [8–10]. We believe that this difference is due to the study design underlying the dataset we used: First, previous studies have restricted task time considerably, while in the essay writing of the Webis-TRC-12 there was no time limit. Second, the required length of the essays is notably longer than in similar studies. Third, the writers of the essays in the Webis-TRC-12 were encouraged to reuse text from search results without originality requirements. These factors likely encouraged authors to copy-and-paste from search results, potentially editing the text later.

In a previous study, Liu and Belkin [8] observed that users kept their search result documents open while moving back and forth between reading documents and writing text. In their scenario, increased usefulness thus comes with increased dwell time. In the case of Webis-TRC-12 instead, many writers first selected the useful pieces from some search result, pasted them into their essay, and modified them later [15]. Thus, the actual dwell time on useful search results is lower in the Webis-TRC-12. Furthermore, the selection of useful text fragments likely resembles relevance assessments. It has been shown that it takes less time to identify a relevant document compared to a borderline case [21, 22]; essay writers likely needed less time to identify useful text passages in search results containing plenty of useful information, compared to documents with less such information. This can also further explain the negative association between dwell time and usefulness in the scenario of our study.

Our study places users into a simulated web-search setting using the ClueWeb corpus, but we believe our results regarding information use for writing tasks

apply also to the wider Digital Library context; while our experimental setting excludes access modalities such as a catalog or a classification system, and limits writers to retrieving sources using a full-text keyword search interface, the latter is clearly a major mode of access in modern, large digital libraries [23]. That said, it is worth exploring if the correlations we observe hold also for other modes of digital library search.

We further believe that our results can be generalized to arbitrary writing tasks of long texts: In essay writing, it is likely that querying and result examination behavior is similar regardless of originality requirements, while text editing will vary by originality. An interesting future research question is how search and text editing contribute to document usefulness in the form of information use, in the presence of stricter originality requirements. In the paragraphs above, we conjecture processes that could explain the associations between the predictors and the usefulness measures. While our regression models do not allow us to test these conjectures, such an analysis could form a promising future direction.

In both our regression models, three predictors cover 91–95% of the explained variation. In both cases, one of these is a query variable, one is a click variable, and one is a text editing variable. Thus, all three variable types are required for an accurate prediction of usefulness based on information usage. Click variables have the strongest effect on usefulness compared to query or text editing variables. However, it is essential to include also the latter ones in the models, as they cover a notable proportion of variation in usefulness. Consequently, personalization in real-world retrieval systems based on information use should include the major factors in these three variable groups, due to their strong effects.

We consider retrieval personalization based on actual information use a promising proposition: the query and click variables we measure are already logged in standard search logs. Beyond that, modern web search engines tend to be operated by companies that also offer writing support tools, and may very well be able to measure text editing variables, as well. By showing that writers' aggregate retrieval success can be predicted by a simple model consisting of three variables, our study takes a first tentative step in this direction. Directly predicting the utility of individual candidate documents for a particular writing task will be important future work, in order to apply this idea in practice.

References

1. Belkin, N., Cole, M., Liu, J.: A model for evaluating interactive information retrieval. In: SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston. (2009)
2. Hersh, W.: Relevance and retrieval evaluation: Perspectives from medicine. J. Am. Soc. Inf. Sci. **45**(3) (April 1994) 201–206
3. Järvelin, K., Vakkari, P., Arvola, P., Baskaya, F., Järvelin, A., Kekäläinen, J., Keskustalo, H., Kumpulainen, S., Saastamoinen, M., Savolainen, R., Sormunen, E.: Task-based information interaction evaluation: The viewpoint of program theory. ACM Trans. Inf. Syst. **33**(1) (March 2015) 3:1–3:30

4. Cooper, W.S.: On selecting a measure of retrieval effectiveness. JASIST **24**(2) (1973) 87–100
5. Vakkari, P.: Task based information searching. ARIST (1) (2003) 413–464
6. Yilmaz, E., Verma, M., Craswell, N., Radlinski, F., Bailey, P.: Relevance and effort: an analysis of document utility. In: Proc. CIKM'14, ACM (2014) 91–100
7. Kelly, D., Belkin, N.J.: Display time as implicit feedback: understanding task effects. In: Proc. SIGIR'04, ACM (2004) 377–384
8. Liu, J., Belkin, N.J.: Personalizing information retrieval for multi-session tasks: the roles of task stage and task type. In: Proc. SIGIR'10, ACM (2010) 26–33
9. Liu, C., Belkin, N., Cole, M.: Personalization of search results using interaction behaviors in search sessions. In: Proc. SIGIR'12, ACM (2012) 205–214
10. Mao, J., Liu, Y., Luan, H., Zhang, M., Ma, S., Luo, H., Zhang, Y.: Understanding and predicting usefulness judgment in web search. In: Proc. SIGIR'17, New York, NY, USA, ACM (2017) 1169–1172
11. Serola, S., Vakkari, P.: The anticipated and assessed contribution of information types in references retrieved for preparing a research proposal. JASIST **56**(4) (2005) 373–381
12. Ahn, J.w., Brusilovsky, P., He, D., Grady, J., Li, Q.: Personalized web exploration with task models. In: Proc. WWW'08, New York, NY, USA, ACM (2008) 1–10
13. He, D., Brusilovsky, P., Ahn, J., Grady, J., Farzan, R., Peng, Y., Yang, Y., Rogati, M.: An evaluation of adaptive filtering in the context of realistic task-based information exploration. IP & M **44**(2) (2008) 511–533
14. Sakai, T., Dou, Z.: Summaries, ranked retrieval and sessions: A unified framework for information access evaluation. In: Proc. SIGIR'13, ACM (2013) 473–482
15. Potthast, M., Hagen, M., Völske, M., Stein, B.: Crowdsourcing Interaction Logs to Understand Text Reuse from the Web. In: Proc. ACL'13, Association for Computational Linguistics (August 2013) 1212–1221
16. Potthast, M., Hagen, M., Stein, B., Graßegger, M., Michel, M., Tippmann, M., Welsch, C.: ChatNoir: A Search Engine for the ClueWeb09 Corpus. In: Proc. SIGIR'12, ACM (August 2012) 1004
17. Hagen, M., Potthast, M., Stein, B.: Source Retrieval for Plagiarism Detection from Large Web Corpora: Recent Approaches. In: CLEF'15 Evaluation Labs, CLEF and CEUR-WS.org (September 2015)
18. Hagen, M., Potthast, M., Völske, M., Gomoll, J., Stein, B.: How Writers Search: Analyzing the Search and Writing Logs of Non-fictional Essays. In Kelly, D., Capra, R., Belkin, N., Teevan, J., Vakkari, P., eds.: Proc CHIIR'16, ACM (March 2016) 193–202
19. Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.: Multivariate data analysis. Prentice-Hall, New Jersey (2010)
20. Hassan, A., Jones, R., , Klinkner, K.: Beyond DCG: user behavior as a predictor of a successful search. In: Proc. WSDM'10, ACM (2010) 221–230
21. Gwizdka, J.: Characterizing relevance with eye-tracking measures. In: Proceedings of the 5th Information Interaction in Context Symposium, ACM (2014) 58–67
22. Smucker, M., Jethani, C.: Time to judge relevance as an indicator of assessor error. In: Proc. SIGIR'12, ACM (2012) 1153–1154
23. Weigl, D.M., Page, K.R., Organisciak, P., Downie, J.S.: Information-seeking in large-scale digital libraries: Strategies for scholarly workset creation. In: Proc. JCDL'17. (June 2017) 1–4