This is the accepted manuscript of the article, which has been published in Information Processing & Management, 2019, 56(3), 879-894. https://doi.org/10.1016/j.ipm.2019.02.001

Modeling the Usefulness of Search Results as Measured by Information Use

Pertti Vakkari[a,*], Michael Völske[b,*], Martin Potthast[c], Matthias Hagen[d], Benno Stein[b]

[a] *University of Tampere, FIN-33014, Tampere, Finland*
[b] *Bauhaus-Universität Weimar, 99423 Weimar, Germany*
[c] *Leipzig University, 04109 Leipzig, Germany*
[d] *Martin-Luther-Universität Halle-Wittenberg, 06108 Halle, Germany*

Abstract

The documents retrieved by a web search are useful if the information they contain contributes to some task or information need. To measure search result utility, studies have typically focused on perceived usefulness rather than on actual information use. We investigate the actual usefulness of search results—as indicated by their use as sources in an extensive writing task—and the factors that make a writer successful at retrieving useful sources. Our data comprise 150 essays written by 12 writers whose querying, clicking and writing activities were recorded. By tracking authors' text reuse behavior, we quantify the search results' contribution to the task more accurately than before. We model the overall utility of the search results retrieved throughout the writing process using path analysis, and compare a binary utility model (*Reuse Events*) to one that quantifies a degree of utility (*Reuse Amount*). The *Reuse Events* model has greater explanatory power (63% vs. 48%); in both models, the number of clicks is by far the strongest predictor of useful results—with β-coefficients up to 0.7—while dwell time has a negative effect (β between -0.14 and -0.21). As a conclusion, we propose a new measure of search result usefulness based on a source's contribution to an evolving text. Our findings are valid for tasks where text reuse is allowed, but also have implications on designing indicators of search result usefulness for general writing tasks.

1. Introduction

Although topical relevance has been an established indicator of effectiveness in evaluating information retrieval systems, the worth [4], utility [15, 42], or usefulness [2] of search result documents has been proposed as an alternative

*Corresponding author
Email addresses: pertti.vakkari@uta.fi (Pertti Vakkari), michael.voelske@uni-weimar.de (Michael Völske)

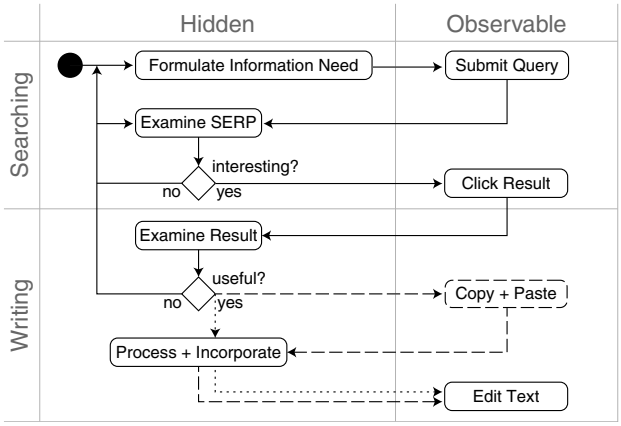


Figure 1: User actions in the search and writing process: our study design involving text reuse (dashed lines) allows for more direct observation of writers' assessments of actual search result usefulness than would be possible without reuse (dotted lines). From [44].

or a supplement since the 1970s [4]. The degree to which retrieved documents help users accomplish the larger tasks they are pursuing with the help of a retrieval system ultimately depends on the information use in the task performance process, be it a decision about what movie to watch or how to write a research proposal. Previous studies on the usefulness of search result documents have not dealt with the utility of information as measured by how it actually contributes towards a task, but focused on users' perceptions of usefulness (e.g., [23, 28, 43]). Various factors in the search process have been used to predict the perceived usefulness of search results (e.g., [26]), but there is a lack of studies analyzing how the querying and clicking behavior of users performing a larger task affects the actual usefulness of retrieved documents.

Some works relate the information search process to task performance, and in the process also touch on information use [16]. Most of these, however, focus on how task features affect search process variables. Only a few studies—discussed in more detail in Section 2—analyze how task performance associates with task outcome (e.g., [28, 43]) or information use during searching (e.g. [1, 14]). Our own study analyses and models the usefulness of search result documents for a writing task as measured by actual information use.

Recently, Vakkari et al. [44] have extended the measurement of search result usefulness from mere user perception to the actual use of information from search results, by way of text reuse: they analyze a dataset of writers working on long essays who used a search engine to retrieve their sources, integrating material deemed useful to the essay through copy and paste, possibly followed by editing to match the overall flow of the text (see Figure 1 for an illustration). Using this approach, Vakkari et al. were able to predict an author's level of success at finding useful sources with high accuracy. In the present work, we adopt the same setting and dataset, and apply it to the analysis of the interactions between searching behavior, and the searcher's success

at finding useful sources. Vakkari et al. [44] analyze how each individual predictor is associated with the usefulness of search result documents and to what extent they jointly predict this usefulness, showing a direct effect between each predictor and the usefulness of search results. The work at hand elaborates on these results by showing how the predictors interact for contributing to the usefulness of search result documents, extending the analyses to include both direct and mediated effects of the predictors on search result usefulness. This allows our study to shed light on the processes that connect search behaviour variables to search result usefulness, and gain new insights on how behavioral signals can yield conclusions on search success or struggling.

1.1. Research Objective

We aim to understand how factors in the search and writing process interact to influence retrieval success, where the latter is measured via text reuse. We believe that such insights will help better understand the level of search satisfaction of users engaged in complex writing tasks, and improve the accuracy with which the usefulness of search results can be quantified—we make a suggestion to this effect in Section 5. Ultimately, if reuse-based usefulness can be predicted accurately at the level of individual documents, this may enable new, specialized ranking signals to better tailor result pages to the needs of writing searchers; since current major search engines tend to be operated by companies that also provide web-based word processing applications, we consider this a feasible end goal.

In this paper, we take a step in this direction by analyzing via path modeling how writers' querying and clicking behavior is associated with information use as measured by (a) *Reuse Amount*, i.e., the number of words from result documents that are reused in an essay, and (b) *Reuse Events*, i.e., the number of clicks that result in any information reuse. From these models, we assess to what extent query, click, and text editing variables predict the overall usefulness of the search results retrieved throughout the writing process, the contribution of each predictor to search result usefulness, and the difference between the two aforementioned notions of usefulness.

1.2. Contributions

In pursuit of the research objectives outlined above, this paper makes the following contributions:

1. We develop a usefulness model as indicated by information use in the context of essay writing, parameterized by query and click variables.
2. We show that for the task in question, the number of clicks is by far the strongest predictor of result usefulness while increasing dwell time predicts decreasing usefulness, and we reconcile the latter finding with contradictory results from the literature.
3. We show that the simpler *Reuse Events* model has a greater explanatory power compared to the *Reuse Amount* model based on the number of words obtained from search results for an essay.

4. We propose a new measure of search result usefulness, based on the contribution of a search result as a source to the text representing a task.

In what follows, Section 2 presents a survey of previous studies modeling and predicting search result utility, concluding that, for the most part, the focus has been on modeling subjective assessments of utility, while dwell time was the most prominent predictor. We then outline the data acquisition process, result utility measures, and modeling approach for our own study in Section 3, pointing out key departures from previous work in the process. Section 4 details the results of our analysis and some of their immediate implications, in particular, on the identification of struggling web search users. Section 5 discusses the wider-ranging implications and limitations of our work.

2. Related Work

Search result usefulness is typically understood as the information in retrieved documents being used in some way to advance a larger task that triggered information searching [2, 4, 15, 42]. This implies that subjects scan and read a clicked document to assimilate information for immediate use, or for extracting appropriate information items for later use. It has been proposed that users first scan a document to make an initial assessment of whether it contains useful content, and if the assessment is positive, they continue reading to obtain applicable pieces of information [46]. Thus, the actual usefulness of search results means that users obtain information, which they then use to advance the task at hand—for example, this can involve finding a source for writing a text. Actual and perceived usefulness are closely related: perceived usefulness is a necessary, but not a sufficient condition for actual usefulness.

2.1. Studies Based on Perceived Usefulness

Although actual information use is the ultimate criterion for the usefulness of search results, almost all studies on this topic have measured the perceived usefulness of information as judged by real users or expert assessors. These studies can be roughly divided into two groups: those studying (and in some cases, predicting) the perceived usefulness of search results retrieved for some task, and those comparing relevance and usefulness assessments.

Exploring the Usefulness of Search Results. Kelly and Belkin [22] analyzed the association between document dwell time and usefulness as measured by subjects' ratings of the documents they retrieved to complete their daily information-seeking tasks. The results showed neither a clear relationship between dwell time and document usefulness, nor interaction effects between task, usefulness, and dwell time. Dwell time did not differentiate document usefulness in various tasks over all participants. Kelly [21] later elaborated that with increasing usefulness, dwell time increased for some participants but decreased for others.

Kellar et al. [20] examined the time spent reading relevant and non-relevant search results for three web search tasks: assessing the relevance of documents,

and answering simple or complex questions. In the relevance assessment task, dwell time on relevant and non-relevant documents did not differ, while question answering led to more dwell time on relevant documents. When users had to find a specific piece of information, they spent more time on the article containing that information. This became more pronounced as the complexity of the task increased. Thus, time use as implicit relevance indicator seems to be more useful when the search task requires demanding information processing.

Liu and Belkin [27] investigated whether time spent on a search result predicts its perceived usefulness for writing a feature article over three sessions, and found that the longer a page was displayed in a session, the higher its usefulness was ultimately rated. Users often moved back and forth between reading documents and writing reports. Documents with longer dwell times were more likely to be useful. The first dwell time in a document—referred to as decision time—was found to be an indicator of whether a document was considered useful. Decision time was not linearly associated with usefulness.

Liu et al. [26] modeled user search behavior in lab experiments in order to predict search result usefulness. Their participants rated the usefulness of each page saved for various information gathering tasks. The authors used binary recursive partitioning to identify the most important predictors for usefulness; dwell time was the most important variable in the model, followed by time to the first click, and the number of visits to a page. Long dwell time, more than one visit to a page, and a short time to the first click predicted usefulness. A later study by Liu and Belkin [29] found interaction effects between dwell time and task stage, but the strong correlation between long dwell time and perceived usefulness remained.

Mao et al. [32] modeled the usefulness of search results for answering short questions by content, context, and behavioral factors. Of these, behavioral factors were the most important to determining usefulness judgments, followed by content and context factors. The longer the dwell time on search results, the higher their similarity to the answer, and the fewer previous results visited, the higher the perceived usefulness of search results.

Kim et al. [24] modeled click dwell time to predict click-level satisfaction. They collected click instances with query and click attributes like the type of query or the reading difficulty of a clicked page. Human assessors were asked to review the search sessions—comprising queries and clicks—and to rate the satisfaction associated with every observed click. While not explicitly discussed by the authors, click satisfaction was defined in terms of the extent to which a clicked document corresponds to a given query, i.e., as topical relevance. The results showed that satisfied clicks had longer dwell times, which varied by the query's click attributes, so that, e.g., more dwell time was required for pages with high reading difficulty. Similarly, Yilmaz et al. [46] identified the amount of effort required to find relevant information in a search result as important to the utility of that result to a real user

Comparing Assessments of Relevance and Usefulness. Interest in usefulness as an evaluation measure has led to studies comparing users' usefulness judgments

to expert assessors' topical relevance and usefulness judgments. Kim et al. [23] elicited binary usefulness assessments of results from users searching the web, which were compared to relevance assessments of the same results by trained assessors. With decreasing relevance, judges classified an increasing proportion of results inconsistently with users.

Mao et al. [31] made observations consistent with the above: their users assessed the usefulness of documents retrieved for twelve search tasks, while expert judges assessed both relevance and usefulness. The correlation between users' usefulness assessments and judges' relevance assessments was low, with each variable explaining only about 10% of the others' variation, as was the correlation between users' and judges' usefulness assessments.

Jiang et al. [17] obtained topical relevance and usefulness assessments of search results clicked by users for various types of search tasks. They found a high correlation between in situ usefulness assessments on the one hand, and post-session topical relevance and usefulness assessments on the other. Click dwell time was the strongest predictor of post-session usefulness.

In all, most of the studies on the usefulness of search results focus on dwell time as the predictor of usefulness. Only few studies analyze the association of querying with the usefulness of search results (e.g. [25]). Most studies find dwell time to be a significant predictor of search result usefulness: the longer the dwell time, the more useful the document [17, 23, 26, 27]. This corresponds to findings that dwell time on topically relevant documents is a good predictor of click satisfaction (e.g. [24]). In addition, studies show that the consistency between the assessments of topical relevance and usefulness of documents is relatively low, regardless of whether the assessors are users or a combination of users and expert judges [23, 31]. By contrast, Jiang et al. [17] find a higher consistency of topicality and usefulness assessments.

2.2. Studies Based on Actual Usefulness

By comparison, only few studies assess the actual usefulness of information gathered from a retrieval system: Ahn et al. [1] and He et al. [14] analyzed how search systems support searching, collecting and organizing useful notes that provide answers to questions in task scenarios. In their studies, human annotators assessed the utility of passages for each task scenario, and task models were used for personalization. Sakai and Dou [38] proposed an evaluation metric called U-measure based on trailtext—the concatenation of all the texts read by the user during a search session. U-measure quantifies the usefulness of a search session based on the assumption that trailtext is used for the task that generated the search session.

As far as we know, however, no studies have investigated how the actual usefulness of the obtained information is associated with search process variables—with the exception of the work by Vakkari et al. [44] discussed in Section 1. While some have explored how the search process is associated with the task outcome, such as the quality of an essay [28, 43, 45], or the knowledge gain in a

Table 1: TREC Web Track 2009 Topic 1 (top) and the derived essay writing prompt (bottom); reproduced from [36].

Query. obama family tree *Description.* Find information on President Barack Obama's family history, including genealogy, national origins, places and dates of birth, etc. *Sub-topic 1.* Find the TIME magazine photo essay "Barack Obama's Family Tree." *Sub-topic 2.* Where did Barack Obama's parents and grandparents come from? *Sub-topic 3.* Find biographical information on Barack Obama's mother.
Obama's family. Write about President Barack Obama's family history, including genealogy, national origins, places and dates of birth, etc. Where did Barack Obama's parents and grandparents come from? Also include a brief biography of Obama's mother.

search-as-learning setting [7, 47], they have not explicitly dealt with the actual use of information.

3. Quantifying Search Result Usefulness

In the following, we outline our data collection procedure, define the measures of search result usefulness for our study, and describe the relevant variables, as well as the modeling approach we use to study the association between search behavior and search result usefulness.

3.1. Dataset

Our analysis is based on the Webis Text Reuse Corpus 2012 (Webis-TRC-12),[1] which comprises 150 essays written by 12 writers [37, 36] during a crowdsourcing study conducted in 2012. Each essay in the corpus was written about one of the 150 topics from the TREC Web Tracks 2009–2011. Since the TREC topics were originally designed for searching for information, and not amenable for the intended purpose as is, they were rephrased so that they ask for writing an essay instead. Table 1 shows an example topic: as Potthast et al. [36] point out, some of the TREC subtopics were omitted from the writing prompts if they were deemed too specific, as was the case with Sub-topic 1 here.

Study participants were instructed to retrieve sources from a static web corpus, and to reuse them to write their essays. There was no time limit for accomplishing the task. Writers' interactions with the search engine, as well as revisions to their essays were recorded in a fine-grained manner. The dataset's constructors had several research tasks in mind to which the corpus contributes, including the study of search behavior in complex, exploratory information retrieval tasks [10], the study of plagiarism and retrieving its sources [9], and the study of writing behavior when reusing text and paraphrasing it [36].

For the purpose of constructing the corpus, a static web search environment was built, consisting of the ChatNoir search engine [35][2] that indexes the

[1] https://webis.de/data/webis-trc-12.html
[2] https://chatnoir.eu

Table 2: Number of essays by author in the study, and means of basic search and writing behavior; non-native English speakers are marked with an asterisk.

Author	Essays	Average number of		
		Queries	Clicks	Words
u002	33	186.4	174.4	4869.1
u017	23	108.3	74.0	5003.4
u018	20	67.8	48.5	5143.4
u005	18	50.4	111.6	4367.6
u007*	12	59.2	117.2	7197.9
u021*	12	66.0	219.7	4979.9
u024	11	23.6	97.7	4987.2
u020	10	53.3	68.9	4801.2
u006	7	57.1	22.0	3693.3
u001*	2	23.5	146.5	4525.5
u014	1	15.0	65.0	4830.0
u025	1	68.0	27.0	5007.0

ClueWeb09, and a web emulator that, for a given address, returns the corresponding web page from the crawl [36]. Hyperlinks found on returned pages were rewritten on-the-fly to refer back to the web emulator instead of the live web. This way, the ClueWeb could be searched and browsed without relying on web pages still being accessible at their original address, while the search engine and the web emulator kept user-specific access logs. The writers used an online rich text editor which logged revisions to the essays by storing the current version of a text whenever its writer paused for more than 300 ms. The search log, the browsing log, and the writing log represent a complete interaction log from being prompted with a writing task to the finished essay.

Having prepared the 150 topics, editor and search environment, twelve writers were hired via an ad placed on the crowdsourcing platform Upwork (still named oDesk at the time of the study); hourly rates were negotiated individually, the median being 11 USD. The writers were instructed to choose an available topic to write about and to write an essay of at least 5000 words length, using only the supplied search engine and rich text editor. Writers were explicitly asked to reuse (and optionally modify) passages from search results. Based on questionnaires that the writers completed after finishing their tasks, Potthast et al. [36] provide detailed demographic data: on average, the writers are middle-aged, well-educated native English speakers, with about 8 years of professional writing experience, and daily web search engine users. Table 2 shows the distribution of essays by writers, along with basic statistics of searching and writing behavior; writers who did not list English among their native languages are marked with an asterisk.

The table shows a fair amount of variation in basic search behavior, and an uneven distribution of topics among writers. Due to the small number of writers, this may cause the distributions of some behavioral variables to be biased. In order to find out the extent of such bias, we compared the means of

those factors which were significantly associated with search result usefulness (see Table 3); an ANOVA indicated that there were significant differences between the writers both in query, click and usefulness variables ($p<.001$). A post hoc analysis (Dunnett C) showed that the writers u002 and u005, in particular, differ significantly from the others. However, these two writers also differ significantly from each other in almost all variables analyzed, and thus any biasing effect of their search behavior will tend to cancel out in subsequent modeling efforts.

3.2. Operationalizing the Usefulness of Search Results

In its most general interpretation, usefulness means that information is obtained from a search result document to contribute to a favorable task outcome. Here, we focus only on cases where information is directly extracted from a search result, not where it is first assimilated and transformed through the human mind. In the context of our study, information is useful if it is extracted from a source and placed into an evolving information object: text is copied from a retrieved document and pasted into the essay (see Figure 1).

The writers generally adopted a workflow of copying text from the sources they retrieved, and then rewriting it to fit the flow of their essay [10]. Since they were encouraged to reuse text at the outset, the writers had a natural incentive to select and copy source passages that require little rewriting to fit and enhance the evolving essay text. This implies that the act of copying and pasting a passage of text will reflect the usefulness of that passage's source quite accurately, even more so than in a situation where source text had to be rephrased or modified to a greater extent. In that sense, information reuse provides a reliable signal for information usefulness that covers a proper subset of the cases in which search results are perceived as useful (cf. Section 2.1): whenever searchers reuse part of a document, that document will have been perceived as useful; there may still be cases where perceived usefulness does not result in reuse (discussed further with regard to the limitations of our study in Section 5.4).

In the following, we explore a quantitative and a binary notion of an individual document's usefulness. Each gives rise to an operationalization for the usefulness of the search results retrieved throughout the writing process:

Reuse Amount. We consider the usefulness of a search result as the amount of text that it contributes to the essay. We operationalize this usefulness notion as a dependent variable by counting the number of words the user reuses from clicked result documents over the entire writing process.

Reuse Events. We consider search results as either useful or not based on whether they contribute any text to the essay. To operationalize this usefulness notion as a dependent variable, we count the number of times text is extracted from result documents over the entire writing process.

These measures are limited in that they neither account for the subsequent editing of pasted information, nor for the importance of the obtained text

Table 3: Means and standard deviations of study variables (n=150; dependent variables marked *).

Variable	Mean	Stddev
Querying		
Queries	46.9	42.2
Unique queries	24.5	17.6
Anchor queries	5.7	6.5
Search sessions	7.1	4.0
Unique terms from SERPs/results	51.3	45.2
Querying time (sec)	2448	2727
Clicking & Result Examination		
Clicks	113.0	81.1
Dwell time (sec)	4877	3969
Useful clicks	32.5	25.7
Writing		
Pastes*	28.0	21.4
Words pasted*	8645	5339
Revisions	2826	1422
Writing time (sec)	21818	13577
Words in essay	4987	1283

passages, instead using the amount of obtained information as a proxy. While the amount and importance of information are likely not linearly related, our presupposition that an increase in the amount of pasted text directly reflects usefulness resembles typical presuppositions made in information retrieval research: for example, Sakai and Dou [38] suppose that the value of a relevant information unit decays linearly with the amount of text the user has to read.

3.3. Independent and Dependent Variables

In order to analyze the associations between the writers' search behavior and the usefulness of search results, we derive 14 basic variables for each of the 150 essays. For each variable, the unit of observation is a single essay, across its entire writing and material gathering process.

Table 3 describes the means and standard deviations of the variables: We measured the number of unique queries and anchor queries submitted while working on the essay, where the latter are queries re-submitted occasionally in order to keep track of the main theme of the task. We subdivided the writers' work into physical sessions whenever a break of 30 minutes or more occurred; search sessions refer to sessions during which a writer submitted at least one query. We recorded the number of unique query terms submitted to the search engine after they occurred in either a SERP snippet or a result document. In addition to the total number of clicks on search results, we recorded the number of "useful clicks," i.e. those that result in at least one text passage being copied and pasted from the document.

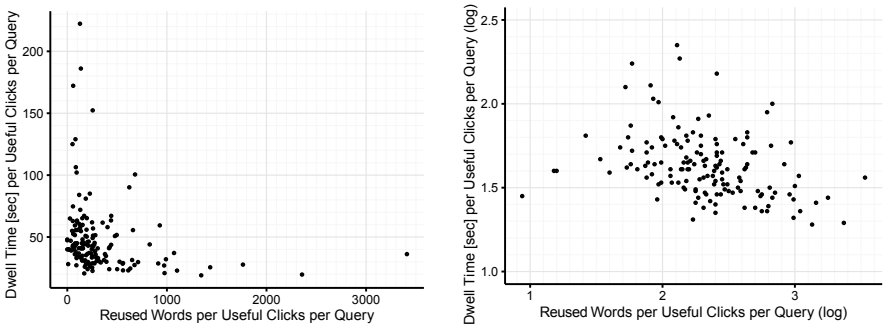
Further variables measure the number of text passages copied from a document ("pastes"), the number of words copied and pasted in this manner, and the time devoted to reading the clicked documents (i.e. dwell time). As mentioned above, a revision refers to a new version of the essay recorded whenever the writer stops typing for more than 300 ms—the number of revisions is thus a measure for the amount of work the writer invested in an essay. For the time variables mentioned in the table, we follow Hagen et al. [10] in computing intervals from the timestamps in the raw interaction log; it should be noted that "Querying time" encompasses the time spent formulating queries and examining the search result pages—this cannot be further distinguished—but is distinct from the time spent examining result documents or writing text.

For an average essay, writers spent 41 minutes formulating queries and browsing SERPs, 1 hour and 21 minutes examining search result documents, and 6 hours and four minutes writing. While they worked, writers pasted about 8645 words across 28 paste events, averaging 309 words per paste. On average, the work on an essay was spread across 7.1 search sessions and 5 writing sessions, which could be distributed over several days. Although these figures may seem large, one has to recall that the task was to search information for writing long essays of 5000 words. There were 7.1 search sessions on average, each lasting on average 5.7 minutes, with 6.6 queries per search session. In an experimental study by Jiang et al. [18], participants submitted 6.2 queries for an exploratory search task during a search session of 10 minutes, while in a log analysis Hassan et al. [13] noted that an average exploratory search session consisted of 4.5 unique queries. Thus, the query profile in our data does not differ much from these two studies.

From the variables in Table 3 we derive an initial set of nine candidate predictors as ratios between the basic variables, such that they cover all the *Querying* and *Clicking* variables from the table. Following Vakkari et al. [44], we infer these candidate variables and normalize them by the number of queries to render different essays comparable:

1. the number of seconds spent querying per query,
2. the number of clicks per query,
3. the proportion of useful clicks out of all clicks per query,
4. the number of unique query terms per query,
5. the dwell time on search result documents per click per query,
6. the proportion of unique query terms out of all query terms per query,
7. the number of unique query terms obtained from search results per query,
8. the proportion of unique queries out of all queries,
9. the proportion of anchor queries out of all queries.

Similarly, we form two measures of search result usefulness according to the usefulness notions discussed in Section 3.2 as dependent variables: (a) the number of words pasted per useful click per query to quantify *Reuse Amount*, and (b) the number of useful clicks per query to quantify *Reuse Events*.



(a) Original scale (b) Logarithmic scale

Figure 2: Scatter plots of study variables (n=150).

3.4. Path Analysis

In order to investigate the connections between user behavior and the dependent variables, we employ path analysis, a special case of structural equation modeling which facilitates the discovery of direct and mediated effects of the search process on the usefulness of the search results. A path model identifies the effect of each independent variable on the variance of a dependent variable [11]. Thus, the model indicates the relative effect of each variable on other variables. The relations among the variables in the model are assumed to be linear and without interactions; a path coefficient (β) indicates the direct effect of a variable hypothesized as a cause on a variable taken as an effect. Path coefficients are standardized regression coefficients obtained through ordinary regression analysis. Typically, more than one regression analysis is called for: at each stage, a variable taken as a dependent is regressed on the variables upon which it is assumed to depend.

The aforementioned linearity assumption held for all variables under consideration, with one notable exception: as evidenced by a high Spearman's ρ combined with a low Pearson correlation, the associations between dwell time and the *Reuse Amount* measure (number of reused words per useful click per query, ρ=-.42, r=-.23)—as well as between dwell time and the *Reuse Events* measure (number of useful clicks per query, ρ=-.52, r=-.26)—are non-linear. Since path analysis requires linearity, we apply logarithmic transformations using a base of 10, which achieves the desired effect: the non-linear association between dwell time and the *Reuse Amount* measure disappears, while it decreases notably between dwell time and the *Reuse Events* measure; at the same time, the transformations enhance the linear correlation of dwell time with *Reuse Amount* to r=-.40, and with *Reuse Events* to r=-.43. Figures 2a and 2b illustrate the effect, showing scatter plots between the dwell time per useful click per query, and the number of reused words per useful click per query, before and after the transformation.

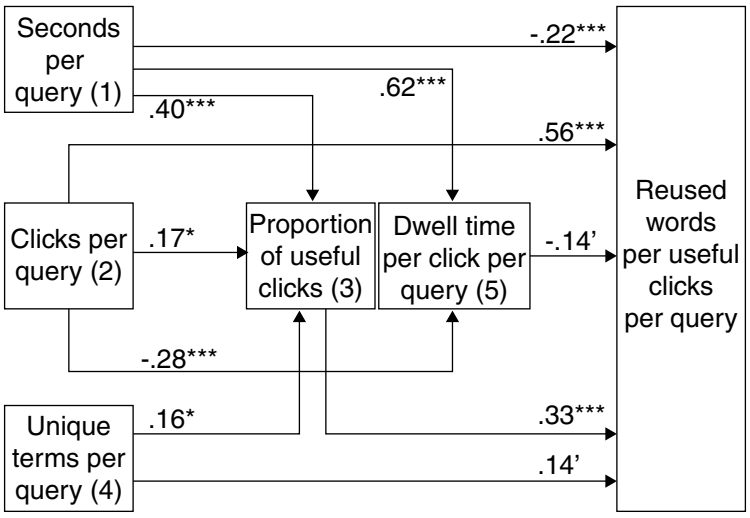


Figure 3: A path model for the *Reuse Amount* notion of search result usefulness, as measured by the number of *reused words per useful clicks per query* (n=144). Independent Variables are numbered as in Section 3.3. Significant paths are shown along with their β-coefficients, annotated with significance levels.[3]

4. Path Models for Search Result Usefulness

We construct a path model for each of the two notions of search result usefulness described in Section 3, beginning with a measure for *Reuse Amount* as the dependent variable. The resulting model permits new insights into the search and writing process of our study participants: in a nutshell, authors who struggle with their task tend to invest more time and require more work sessions to complete their essays; struggling across different phases of the writing process tends to accumulate. However, we find that a second path model with *Reuse Events* as the dependent variable has a greater explanatory power.

4.1. Modeling the Amount of Reuse

We begin by constructing a path model for search result usefulness as measured by the number of words reused in—i.e., pasted into—the essay. From the candidate predictors listed in Section 3.4, we select five by running a regression analysis, retaining those for the path model that associate notably ($p<.10$) with the number of reused words per useful click per query.[4] For the regression analyses, we omitted two outlier essays outside three standard deviations in

[3] Key: ' = ($p<0.10$), * = ($p<0.05$) , ** = ($p<0.01$) , *** = ($p<0.001$)

[4] The following candidate predictors were removed due to their non-significant contribution: the number of unique query terms obtained from results (4), the proportion of unique query terms per query (6), the proportion of unique queries out of all queries (8), and the proportion of anchor queries out of all queries (9).

standardized regression residuals, and four essays with missing values, yielding n=144 essays. Figure 3 shows the resulting path model, which is significant (R=.70; R^2=.49; AdjR^2=.48; F=45.2; p<.001). The model explains 48% of the variation in the number of words from search results.

The predictors seconds per query (β=-.22***), number of clicks per query (β=.56***), and the proportion of useful clicks out of all clicks (β=.33***) have significant direct effects on the amount of useful information reused from search results, while the number of unique terms per query (β=.14'), and dwell time per click per query (β=-.14') have notable effects. With the exception of dwell time, click variables have a considerably stronger contribution to the number of words obtained from search result documents than query variables. Writers extract more words from search results when they spend less time formulating queries (β=-.22***), use more unique terms per query (β=.14'), click more (β=.56***), have proportionally more useful clicks (β=.33***), and spend less time reading search results (β=-.14').

The more time writers spend querying, the fewer words they paste from search results (β=-.22***). Querying time is also strongly associated with spending more time examining search results (β=.62***), which further decreases the number of reused words (β=-.14'). Conversely, querying time contributes positively to the number of reused words by way of the proportion of useful clicks: the more time writers spend querying, the larger the proportion of clicks that are useful (β=.40***), which increases the number of reused words (β=.33***). The other query variable, the number of unique terms per query, directly increases the number of reused words (β=.14'), but also positively affects the proportion of useful clicks (β=.16*), which indirectly further increases the number of reused words (β=.33***).

The number of clicks is not associated with the query variables in the model. Clicks have, however, a strong direct positive effect on the number of reused words (β=.56***). In addition, two indirect paths mediate the contribution of clicks: First, the proportion of useful clicks mediates (β=.17*) the effect of clicking by increasing the number of reused words (β=.33***). Second, the number of clicks decreases the time allocated to reading search results (β=-.28***), which in turn increases the number of reused words (β=-.14').

4.2. Does Time Use Signal Struggling or Success in a Search?

The path model hints at an accumulating mechanism affecting the variation in the number of reused words. Depending on the point of view of analysis, the mechanism either tends to decrease or to increase the *Reuse Amount*.

Looking at the decreasing paths, the more time writers spend querying, the less text they obtain from search results. The increase in querying time also increases the time devoted to examining search results, which decreases the number of reused words. Since a positive association along some path in the model implies also that a *decrease* in the independent variable decreases the dependent variable, we find additional mechanisms that can reduce the number of reused words: A decrease in the number of clicks directly decreases the number of reused words, while it increases dwell time, and decreases the

proportion of useful clicks, which both reduce the number of reused words. Finally, the fewer unique query terms are included in individual queries, the lower the proportion of useful clicks, and consequently, the fewer words are reused from search results.

Naturally, the converse applies to the independent variables negatively associated with the dependent variable: Decreasing dwell time or querying time increases the number of reused words. An increase in the number of clicks increases the proportion of useful clicks and reduces dwell time, both increasing the number of reused words. Also, an increase in the number of unique terms per query enhances the proportion of useful clicks, leading to an increase in the number of reused words.

This mechanism can be interpreted in two different ways: the lower amount of reused text either reflects selectivity, or difficulty identifying useful information. Under the first interpretation, writers invest a considerable amount of time carefully formulating queries, leading to high quality result lists; similarly, they may meticulously analyze and locate useful pieces of information in search results. Either behavior (or both simultaneously) would lead to a smaller amount of reused text, yet easier to be incorporated into the developing essay. Greater selectivity may thus be associated with fewer clicks, more time examining results, and a lower proportion of useful clicks, and would lead to the reduced amount of reused text.

Under the second interpretation, writers have difficulties to formulate queries and to identify useful information in search results, which leads to an increase in query formulation time and dwell time, and a decrease in the amount of text reused. The reused text requires additional editing effort to match the evolving essay text. A reason for difficulties in finding useful information may be the low quality of the result lists produced by low quality queries. Consequently, the documents on the list likely contain information with varying degrees of utility, among which it is more laborious to identify useful search results and accurate information within results. This may be reflected in the decreasing number of clicks, a decreasing proportion of which is useful. Conversely, writers who are able to formulate good queries in a short time make a higher number of useful clicks, and quickly identify useful information to integrate into the evolving structure of the text.

Since both hypotheses are plausible, we test them by their consequences: If time spent querying yields high quality queries, it should be associated with the number of terms per query obtained from the search results. These terms have been shown to be especially pertinent query terms, positively associated with result list quality (e.g., [5, 40]). A check in the data shows a significant negative association between these two variables (β=-.21; p=.014); increasing query formulation time decreases the number of query terms obtained from search results. Thus, time use in querying damages the quality of queries as indicated by the query terms from results.

As suggested previously, editing effort should vary with the amount of information reused from search results when more text needs to be integrated with the evolving essay text. A check in the data shows that the more words

Table 4: Spearman's ρ between number of search sessions, number of writing sessions, and selected variables, with significance levels. Variables appearing in the path models are numbered as in Section 3.3.

Variable	**Sessions**	
	Search	Writing
Seconds Per Query (1)	.25**	.17*
Unique Terms Per Query (4)	-.27***	-.21**
Unique Terms from Results Per Query (7)	-.21*	-.16*
Clicks Per Query (2)	-.24**	.09
Dwell Time Per Click par Query (5)	.18*	.07
Useful Clicks Per Query (3)	-.23**	-.01
Words from a Useful Click Per Query	-.40***	.02
Edits Per Paste	.23**	.29***
Seconds Writing Per Revision	.22**	-.05

were reused from search results, the fewer edits were done per paste (β=-.28; p=.001) and the less time was used per revision (β=-.29; p<.001). This finding supports the hypothesis that reduced time use both in querying and in result examination reflects effortless query formulation and identification of useful information in search results, which increases the amount of text reused from results, and which better matches the accumulating essay text. In addition, the more time was used formulating queries (β=.43; p<.001) and reading useful search results (β=.42; p<.001), the more time was used for writing per revision, also supporting the notion that decreasing time use reflects effortless query formulation and identification of useful information.

In sum, increasing time use in query formulation, reading result documents, and editing pasted text passages, seems to indicate that some writers have problems in all these stages of the search and writing process. This all hints that writers who struggle with query formulation struggle also with finding useful pieces of information in the documents, and consequently, with editing the text to match to the structure of the essay, and vice versa.

4.3. Struggling and Task Sessions

When time use among struggling writers is excessive in various phases of the search and writing process, it can be supposed that their task performance is distributed over more search and writing sessions than among writers who do not struggle. To test this supposition, we examine the correlation between relevant variables and the number of both session types; due to the non-linearity of many associations, we use Spearman's ρ. Table 4 shows that the number of both search and writing sessions are associated with the search and writing effort variables as expected. The number of both session types significantly increases with effort in query formulation (ρ=.25**; ρ=.17*), in examining search results (ρ=.18*), in editing pastes (ρ=.23**; ρ=.29***) and revising text (ρ=.22***). Thus, the more time and effort writers have to invest in succeeding

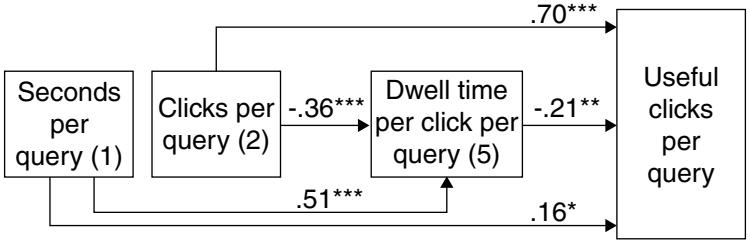


Figure 4: A path model for the *Reuse Events* notion of search result usefulness as measured by the number of *useful clicks per query* (n=150).

in various phases of task performance, the more search and writing sessions this process consists of. An increase in the number of both session types correlates with a decreasing number of unique query terms obtained from result lists (ρ= -.21*; ρ= -.16*), clicks (ρ=-.24**), useful clicks (ρ=-.23***) and words (ρ=-.40***) per click, as well as in a decreasing number of unique query terms per query (ρ=-.27***; ρ= -.21**). It seems that an increasing struggle in query formulation and result examination—likely produced by difficulties in shaping the content of an essay—leads writers to divide the work over more sessions.

As can be expected, the number of search sessions is more strongly associated with the variables reflecting querying and result examination compared to the number of writing sessions: struggling in searching is naturally reflected in an increase in search sessions. A failure to find useful text passages from search results leads to the use of several search sessions for finding material for the essay. A strong association between searching and writing sessions (ρ=.41***) indicates that difficulties in searching are also reflected in the writing process. As stated previously, struggling in various phases of the whole task process accumulates, which increases the number of both search and writing sessions. Conversely, if the various phases of the search and writing process go smoothly and successfully, fewer sessions of both types are needed to finalize the essay.

4.4. Modeling Reuse Events

As an alternative to modeling the usefulness of search result documents by the number of words writers extract for use in their essays, we consider the *Reuse Events* notion of search result usefulness, where a search result either provides any amount of content for the essay or not. We compare the two models with the aim of analyzing to what extent they resemble each other, and to assess to what extent the two usefulness indicators are interchangeable.

Under the *Reuse Events* notion, a search result document is useful if it produces at least one paste event. The dependent variable in the model then is the number of useful clicks per query. To select the independent variables for this path model, we revisit the nine candidate predictors that were under consideration for the previous model, except for the proportion of useful clicks of all clicks. The latter is excluded because, in causal order, it comes after a decision concerning the usefulness of a click. Due to the non-linearity of

association, we again apply a logarithmic transformation to the usefulness variable and the dwell time variable. Since the number of useful clicks and the time spent querying are associated non-linearly as well, we also apply a logarithmic transformation to the time spent querying. We then retained only those variables that are associated at least notably ($p<.10$) with the usefulness of search results under a regression analysis. Five out of eight candidate predictors were excluded due to a non-significant association. The resulting path model consists of the remaining variables depicted in Figure 4. The path model is significant ($R=.79$; $R^2=.63$; $AdjR^2=.63$; $F=79.4$; $p<.001$) and explains 63% of the variation in the number of useful clicks.

All the predictors have significant direct effects on the dependent variable. Among them, the number of clicks plays an essential role in contributing to the number of useful clicks ($\beta=.70^{***}$), compared to the dwell time per click per query ($\beta=-.21^{**}$) and to the seconds devoted to querying ($\beta=.16^{*}$).

The mechanism connecting predictors to the number of useful clicks resembles the pattern in the *Reuse Amount* model. The more time writers spend querying, the more time they also spend reading the clicked documents ($\beta=.51^{***}$), which decreases the usefulness of clicks ($\beta=-.21^{**}$). Clicking increases the number of useful clicks directly ($\beta=.70^{***}$) and also indirectly by reducing the time devoted to examining the clicked documents ($\beta=-.36^{***}$), which increases the usefulness of clicks ($\beta=-.21^{**}$). Thus, those writers who spend lots of time formulating queries also tend to spend a lot of time examining clicked documents, which reduces the number of useful clicks. Abundant clicking reduces the time spent reading clicked documents, which increases the number of useful clicks.

4.5. Comparison of the Two Models

Compared to the more fine-grained *Reuse Amount* model, a bigger proportion of the variance of the dependent variable (63% vs. 48%) is covered by the *Reuse Events* model. The contribution of the strongest predictor in both models also differs: the number of clicks per query accounts for much more variance of the *Reuse Events* model compared to the *Reuse Amount* model. In the *Reuse Events* model, it accounts for .597 of the R^2 change compared to .011 for dwell time and .016 for querying time. The corresponding figures in the *Reuse Amount* model are .395 (number of clicks), .057 (the proportion of useful clicks) and .040 (querying time). Thus, the major factor, the number of clicks, contributes notably stronger to usefulness as indicated by the *Reuse Events* measure.

In addition, the associations between the number of clicks per query and other factors are stronger in the *Reuse Events* model compared to the *Reuse Amount* model: the number of clicks has a notably stronger positive effect on the usefulness of clicks, and a negative effect on the time used for reading documents. By contrast, the associations of querying time to other variables are weaker in the *Reuse Events* model. Thus, it seems that in the *Reuse Events* model, click variables play a greater role for explaining the usefulness of clicks compared to query variables. The more fine-grained *Reuse Amount* indicator seems to respond more sensitively to factors in the search process.

Finally, the correlation between the number of useful clicks per query and the number of pasted words per useful click per query is very strong (β=.77; $p<.001$). The proportion of joint variance between the measures is 59% as indicated by Adjusted R^2, which explains the relatively similar behavior of the two models.

5. Discussion

Previous studies have represented usefulness mostly as perceived by the users, not as the actual use of information. As far as we know, our study is the first attempt to model the usefulness of search results in terms of how the information from search result documents actually contributes to a larger task. The results extend our knowledge about which factors in the search process predict the usefulness of information in search results for essay writing tasks. Our findings have applications in retrieval enhancements based on usefulness prediction, and in the design of indicators for the usefulness of search results.

5.1. Models with Differing Notions of Usefulness

We employed query and click variables to model the number of pasted words from clicked search results per query; the resulting *Reuse Amount* model covers 48% of the variation in the dependent variable. For the sake of comparison, we also modeled the usefulness of search results as a more traditional, binary dependent variable, considering a search result as useful if it produced a paste for the essay, regardless the number of words pasted. This *Reuse Events* model covers 63% of the variation in the usefulness of search results. Thus, the *Reuse Events* model accounts for a considerably bigger proportion of variance compared to the more fine-grained *Reuse Amount* model.

This difference likely reflects the differences in validity between these indicators: the number of pastes evidently reflects the actual usefulness of search results more validly than the number of pasted words. It is plausible that at least a part of the pasted text is assimilated into the essay, implying actual use. In that sense pasting means use. Instead, the number of pasted words reflects "potential use," not taking into account what proportion of the pasted text is retained in the final essay.

The path models for both dependent variables are structured in a relatively similar way, both regarding the set of significant predictors, and their associations. In both models, the number of clicks is clearly the strongest predictor of search result usefulness compared to other predictors like the time spent querying or the dwell time in search results. However, the *Reuse Events* model includes fewer query variables than click variables, and the strength of associations on the part of click variables is stronger. Thus, click variables have a greater role than query variables in predicting the usefulness of search results in the *Reuse Events* model. The more fine-grained *Reuse Amount* model is more sensitive to factors in the search process due to the more extensive coverage of both query and click variables. However, the binary *Reuse Events* usefulness

indicator produces a simpler model with greater explanatory power. In both models, the number of clicks predicts search result usefulness more strongly than all other predictors combined, but this effect is much more pronounced in the binary model.

Our results indicate that if one wishes to have a simple, but effective model or one strong predictor of search result usefulness—in our case, the number of clicks—one should represent usefulness by a binary measure. If one wishes to predict search result usefulness by a model which is somewhat more sensitive also to query variables, then it is recommended to represent usefulness by the number of pasted words. As noted, this sensitivity comes at the cost of a smaller explanatory power.

5.2. Decreasing Dwell Time Predicts Useful Search Results

Both models share much of their predictors and associations, including the following patterns: (1) The shorter the time spent formulating a query, (2) the more clicks per query, and (3) the shorter the dwell time in search results per query, the more useful the search results were. In previous research, the number of clicks has been associated to satisfaction with search. Fox et al. [6] found that a short dwell time on results and several clicks for a query were both indicators of dissatisfaction. Hassan et al. [12] showed that in a successful search, users were nearly twice as likely to click on a search result as in an unsuccessful one. In addition, studies seem to show almost consistently that the more time users spend on a search result, the more valuable it is, regardless of whether it indicates search satisfaction [6], success [12, 24] or the usefulness of results [27, 26]. Diverging from the consensus of previous studies—that long dwell time predicts search result usefulness—our results instead indicate that the shorter the dwell time, the more useful the search results.

This contradiction concerning the impact of click dwell time on usefulness likely results from differences in the characteristics of the tasks for which information was searched and used: previous studies have limited task time considerably—e.g., to 40 minutes in Liu and Belkin [27] for each of the consecutive sessions the task consisted of, and to 20 minutes in Liu et al. [26]—while in our study there was no time limit for composing the essay. Our writers worked on a median essay for 6.5 search sessions and 10 writing sessions, which could be distributed over several days. Also the required length of the essays (5000 words) was notably longer than in similar studies. Our task was much more laborious compared to those in previous work.

It is likely that in a task with a time limit, participants are working with information while reading the search results. Time pressure and the nature of the task require simultaneously examining results and writing based on the examination. This was the case in Liu and Belkin [27], who report that their users often moved back and forth between reading documents and writing reports, and that those documents which had longer dwell times were more likely to be useful. Thus, a search result that provided lots of information for the report was also kept open for a long time in order to immediately extract and edit the useful text for the report.

Our writers had a more extensive essay writing task, for which it was more effective to copy and paste text from the clicked search results for later use, instead of trying to use information immediately during the first reading. The writers first selected useful text fragments, which they later edited to match the essay text. This implies that the writers had to scan and read the documents to decide which parts of the text to copy for later use. The selection of useful text passages likely resembles relevance assessment. Previous research has found that it takes more time to assess the relevance of a partially relevant document compared to a highly relevant one [8]. The results of Smucker and Jethani [39] also indicate that the more difficult it is to determine a document's relevance, the longer the assessment will take. Thus, it is plausible that our writers needed less time to identify and copy relevant text passages from search results containing lots of useful information, compared to search results with less such information. This would explain why decreasing dwell time was associated with increasing usefulness of search results.

5.3. Accumulating Effort in the Search Process Signals Struggling

Both models share a major pattern predicting the usefulness of search results: the more time writers use for query formulation, the more time they spend reading search results; both contribute to fewer useful clicks, as indicated by both dependent variables in our study. This negative pattern is also associated with an increasing effort editing the pasted text for the essay. The pattern can also be represented conversely, reflecting a positive accumulation of search activities leading to useful results.

We showed that increasing time use, both in querying and examining search results, is associated with difficulties in these activities. Increase in time use in both predictors contributed to a smaller amount of pasted text, which was associated with increasing effort in editing the pasted text to match the essay. It seems that some writers were struggling through the whole task process from query formulation via selecting useful material for the essay to editing the pasted material. For others, the process was smoother, requiring less time and effort in querying and examining search results. This implied greater usefulness of the information in search results, which required less effort in editing the text for the essay.

Hassan et al. [13] modeled differences between exploring and struggling behavior in long, topically coherent sessions: Users in exploring sessions are engaged in an open-ended and multi-faceted information-seeking task to foster learning and discovery, whereas in struggling sessions, they are experiencing difficulty locating the required information. In the study, the number of clicks per query and dwell time per query were significantly smaller in struggling sessions; the similarity of queries to the first query decreased more in exploring sessions. In a similar vein, Odijk et al. [34] studied searcher behavior in struggling search tasks, and in particular, what distinguishes struggling searchers that ultimately succeed from those that do not. Here, the querying time, query length, the number of clicks on search results, and the dwell time all were higher for successful searchers.

For our essay writers, a decreasing number of both clicks and unique terms per query was associated with a decreasing usefulness of search results, which is in line with the findings above. However, to the contrary, we found decreasing dwell time to be associated with search result usefulness and success in all phases of the search, not with struggling or failing to extract useful information. This difference is likely due to the characteristics of our task as discussed above.

To rule out whether the observed struggling behaviors result from language difficulties, we split the essays into two groups based on whether they were written by an English native speaker (n=124) or not (n=26). For both groups, we separately compute the 95% confidence intervals for the means of all of the struggling signals mentioned above (time spent on query formulation, time spent reading search results, clicks per query, and unique terms per query). For three out of these measures, the confidence intervals overlap, lending credence to the hypothesis that the author's native language has no influence. The only exception is the number of clicks per query, where we find that non-native speakers perform significantly more clicks ($p < 0.0007$, crossing the **-significance threshold of 0.0025 when applying the Bonferroni correction for 4-fold testing, with Cohen's $d = 0.79$). As such, we do find some evidence of differences between native and non-native speakers in the number of clicks per query, with a large effect size. However, the majority of the struggling signals show no such difference, so that our observations cannot be explained by demographic differences alone.

5.4. Limitations of our Study

Although our study provides essential results on factors in the search process that contribute to the actual utility of search results, it is limited in a few ways. First, the unit of observation is an essay: we average query, click, and text editing variables over the whole writing process, although the writing task was realized in several sessions. In other words, we treated the writing process as a cross-sectional event. It is evident that the variables thus averaged will reduce the variation in the phenomenon under investigation, and thus reduce the strength of associations [11]. We conjecture that taking a session as the unit of observation would create a more valid account of search result usefulness. The significant associations we observed between the number of search sessions and the variables in both models support this notion. It is an open question to what extent the predictors of usefulness vary within and between the sessions.

Second, our models do not account for all possible factors that may conceivably affect retrieval success: for instance, the user's level of domain expertise has been found to not only affect task performance, but also querying and relevance assessment strategies [33, 41], and perception of task difficulty [30]. In addition, to what extent the topics, the collection, or the quality of search engine influence search behavior in our case cannot be controlled. These factors may influence the validity—in particular, the generalizability—of our findings: users may attempt to compensate a poorly functioning search engine, or a barren collection, by increasing their effort like, e.g., through more active querying. Assuming low quality of collection and search engine, such an increase

in the number of queries, would still likely lead to poor result lists, implying an increase in both dwell time and in the number of inspected SERPs. This would naturally produce different values in a univariate analysis—e.g., variables would have different means compared to a setting where the search engine or the collection or both have higher quality. However, the multivariate techniques we used analyze the relations between variables, and as our previous example shows, an increase in one variable caused by search engine or collection effects likely produces an increase in another variable. Thus, the direction of change is the same, and likely does not change much the strength of associations (beta coefficients) between these variables; we hence don't expect the validity of the results to be affected very much.

Third, there are ways in which search result documents can be actually useful to the searcher's task, and yet not end up being reused. For instance, a document may convince the searcher to exclude a marginally-related subtopic from the essay that she had previously planned to write about; such a search result could improve the focus of subsequent search actions, without ever being observable through our text reuse-based usefulness measures.

Finally, the *Reuse Amount* indicator of usefulness in our study had a limited validity—as discussed in Chapters 3.2 and 5.1: While the number of useful search results (*Reuse Events*) validly reflects the use of information—because the pasted text very likely contributes to the essay in some way—the number of pasted words only reflects potential usefulness of text fragments. As idealizations, both these indicators do not take into account the qualitative aspects of information use, like the importance of the information in contributing to the outcome of the task. These aspects should be explored in studies to come.

5.5. A New Measure for Search Result Usefulness

We propose a *usefulness measure* to better quantify a search result's usefulness as its contribution to the essay text; in the following, we discuss the rationale and desirable properties for such a measure, and propose two variants for its formulation.

Assuming the user's task—e.g., essay writing—can be represented as a collection of words that changes over time [3], the words in the evolving text can be observed at the points in time immediately before (t_1) and after (t_2) a given search result is retrieved and used as a source—in the setting of the present study, t_1 would be the point in time immediately before, and t_2 immediately after copy-pasting from a search result. We represent the content of the essay at time t by a multiset (V_t, m_t), where V_t is the underlying set of words (vocabulary), and m_t the multiplicity function mapping each word to the number of times it occurs in the essay. For a source introduced between times t_1 and t_2, the set $V_{\text{new}} := \left\{ w \in V_{t_2} \mid m_{t_2}(w) > m_{t_1}(w) \right\}$ comprises the new words introduced; we assume those words to reflect a given source's contribution.

Over the course of the writing process, several different search results may contribute a particular word w to the text. In that case, the importance of a given result document with respect to w is inversely proportional to the number of other results that contribute this word. This consideration gives rise to

a weighting term—analogous to the inverse document frequency [19]—that diminishes the contribution of frequent words to the usefulness measurement. Further, the usefulness measure should be normalized with respect to the document's length, to account for the increased effort in identifying useful content in long texts. Given the above considerations, we define the usefulness U of a search result d for a text as follows:

$$\mathrm{U}(d, t_1, t_2) = \frac{1}{|d|} \sum_{w \in V_{t_2}} \Big(m_{t_2}(w) - m_{t_1}(w) \Big) \cdot \left(1 - \frac{\log\left(|\{d \in D : w \in d\}|\right)}{\log\left(|D| + 1\right)} \right) \tag{1}$$

Here, $|d|$ refers to the number of words in the search result d, t_1 to the time immediately before its use, and t_2 to the time immediately after; we assume that the change in word count $m_{t_2}(w) - m_{t_1}(w)$ is always nonnegative. The second factor in the sum is the aforementioned weighting term, wherein D refers to the set of all search results used as sources over the course of the writing process; the term in the numerator accounts for the number of sources which contain the word w under consideration.

A notable issue with the formulation of the usefulness measure given in Equation 1 is that it can only be computed post-hoc, i.e., at the end of the writing process when the set of all sources is known. For certain use cases—such as relevance feedback—it may be desirable to compute usefulness while the writing process is still ongoing. Thus, we propose the following, incremental, variant:

$$\mathrm{U}_{\mathrm{inc}}(d, t_1, t_2) = \frac{1}{|d|} \sum_{w \in V_{t_2}} \sum_{i=m_{t_1}(w)+1}^{m_{t_2}(w)} \mathrm{c}(i) \tag{2}$$

Here, the credit assignment function c quantifies the fraction of credit a search result receives for a word w, depending on the number of times the essay contained w previously. It is defined as follows:

$$\mathrm{c}(n) = \frac{1}{\log_2(n+1)^{\alpha}}$$

The parameter $\alpha \geq 0$ controls how quickly the credit assignment curve drops off: for instance, with $\alpha = 2$, a search result introducing a word to the essay for the second time receives 40% of the credit, and less than 10% for a word that already occurs eight times in the essay. On the other hand, with $\alpha = 0$, there are no diminishing returns in credit assignment, and all search results receive full credit for all words they introduce, no matter how many times the introduced words already occur in the essay.

The measures proposed in Equations 1 and 2 reflect the amount of information derived from a search result to the evolving text, and thus indicate the information gain provided to the user. The incremental variant in Equation 2 can be computed on-the-fly while a task session is still ongoing, at the cost

of overestimating the usefulness of sources introduced early in the writing process. While practically applying these measures may be challenging—it requires that the use of information from each source for a given task can be distinguished from the use of information from other sources—we believe this can be addressed with an innovative experimental setup akin to the one we originally employed to collect the essays.

Information provided by our formula about the new terms in the text from a document can be applied as relevance feedback to re-rank search results, or for query reformulation. This procedure can also be used for query diversification by identifying queries that are associated to the new terms, in particular.

5.6. Generalizability to Other Writing Tasks

Our essay writers were encouraged to reuse text if they could to complete their task. The fact that text reuse was allowed makes our task somewhat different from common writing tasks, where the originality of the resulting text is more important. However, we expect the query formulation and result examination strategies to be more or less the same, regardless of whether or not text reuse is allowed: in either case, the searcher will aim to retrieve as much useful information as possible to help complete the writing task. Beyond that, the processes of identifying useful search results and useful passages within them are likely the same, as well. The actual use of information, i.e., scanning and reading a document, selecting information to be used for the writing task, and writing based on the selection, may be different depending on the time limit for the writing task, or the expected length of the text. As stated above, if the task is time-limited, or if the to-be-written document is short, then writers will likely simultaneously examine results and write new text. If there is no time limit, and if the length of the text is extensive, then writers will likely copy and collect text from retrieved documents for later use, regardless of originality requirements. To what extent the copied text fragments are revised and synthesized depends on the originality requirements of the new text.

In sum, it is likely that query formulation and result examination patterns are relatively similar across writing tasks, regardless of originality expectations, while the actual text writing and synthesizing processes differ. Hence, we expect that our models and usefulness indicators are to a greater extent valid not only where text reuse is allowed, but also in cases where more originality is required, on the condition that the created text is extensive and that there is no time limit for the task.

6. Conclusions

To the best of our knowledge, our study is the first of its kind in that it models search result usefulness as indicated by the actual use of information in opened documents, in the context of the search and writing behaviour of authors engaged in an extensive essay writing task. We have operationalized search result usefulness by two different measures, the quantitative *Reuse Amount*

measure based on the amount of text extracted from results, and the binary *Reuse Events* measure based on whether or not a given click contributed any amount of content to the essay, which allowed us to construct two path models to investigate the associations between measures of querying and clicking behavior with these usefulness measures. The *Reuse Events* model is simpler, more valid and has a greater explanatory power compared to the *Reuse Amount* model. Both models show that the number of clicks is by far the strongest predictor of usefulness, while the dwell time on clicks was negatively associated with usefulness. This latter finding, in particular, contradicts previous results on indicators of click satisfaction. We conjecture that this discrepancy results from the fact that the writing task we studied was much more laborious than those in previous research. It may be of interest to further investigate this effect, and to take it into account in future click satisfaction models.

We further revealed a cumulative struggling and a corresponding success pattern throughout the search process from querying through document exploration to text writing. It is an open question which factors are associated with the actual usefulness of clicks in less extensive writing tasks, not to mention other larger tasks.

Finally, based on the aforementioned insights, we derived a new measure of document usefulness to overcome the limitations of the indicators in this study. It is based on the occurrence frequency of new words in the text representing a task, which can be identified in a freshly used source document. This measure indicates the contribution of a source document to the text. Information about the contributing words can be used for personalization.

References

[1] J. Ahn, P. Brusilovsky, D. He, J. Grady, and Q. Li. Personalized web exploration with task models. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 1–10, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2.

[2] N.J. Belkin, M. Cole, and J. Liu. A model for evaluating interactive information retrieval. In *SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston*, 2009.

[3] J. Budzik and K. J. Hammond. User Interactions with Everyday Applications As Context for Just-in-time Information Access. In *Proceedings of the 5th International Conference on Intelligent User Interfaces*, IUI '00, pages 44–51, New York, NY, USA, 2000. ACM. ISBN 978-1-58113-134-5.

[4] W.S. Cooper. On selecting a measure of retrieval effectiveness. *J. Am. Soc. Inf. Sci.*, 24(2):87–100, 1973.

[5] C. Eickhoff, J. Teevan, R. White, , and S. Dumais. Lessons from the journey: a query log analysis of within-session learning. In *Proc. WSDM'14*, pages 223–332. ACM, 2014.

[6] S. Fox, K. Karnawat, M. Mydland, S. Dumais, , and T. White. Evaluating implicit measures to improve web search. *ACM TOIS*, 23(2):147–168, 2005.

[7] Ujwal Gadiraju, Ran Yu, Stefan Dietze, and Peter Holtz. Analyzing Knowledge Gain of Users in Informational Search Sessions on the Web. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, pages 2–11, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-4925-3. doi: 10.1145/3176349.3176381.

[8] J. Gwizdka. Characterizing relevance with eye-tracking measures. In *Proceedings of the 5th Information Interaction in Context Symposium*, pages 58–67. ACM, 2014.

[9] M. Hagen, M. Potthast, and B. Stein. Source Retrieval for Plagiarism Detection from Large Web Corpora: Recent Approaches. In *CLEF'15 Evaluation Labs*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, September 2015.

[10] M. Hagen, M. Potthast, M. Völske, J. Gomoll, and B. Stein. How Writers Search: Analyzing the Search and Writing Logs of Non-fictional Essays. In Diane Kelly, Rob Capra, Nick Belkin, Jaime Teevan, and Pertti Vakkari, editors, *Proc CHIIR'16*, pages 193–202. ACM, March 2016.

[11] J. F. Hair, W. C. Black, B. J. Babin, and R.E Anderson. *Multivariate data analysis*. Prentice-Hall, New Jersey, 2010.

[12] A. Hassan, R. Jones, , and K. Klinkner. Beyond dcg: user behavior as a predictor of a successful search. In *Proc. WSDM'10*, pages 221–230. ACM, 2010.

[13] A. Hassan, R. W. White, S. T. Dumais, and Y-M. Wang. Struggling or exploring?: disambiguating long search sessions. In *Proc. WSDM'14*, pages 53–62. ACM, 2014.

[14] D. He, P. Brusilovsky, J. Ahn, J. Grady, R. Farzan, Y. Peng, Y. Yang, and M. Rogati. An evaluation of adaptive filtering in the context of realistic task-based information exploration. *Information Processing & Management*, 44(2):511–533, 2008. ISSN 0306-4573.

[15] W. Hersh. Relevance and retrieval evaluation: Perspectives from medicine. *Journal of the American Society for Information Science*, 45(3):201–206, 1994. ISSN 1097-4571.

[16] K. Järvelin, P. Vakkari, P. Arvola, F. Baskaya, A. Järvelin, J. Kekäläinen, H. Keskustalo, S. Kumpulainen, M. Saastamoinen, R. Savolainen, and E. Sormunen. Task-based information interaction evaluation: The viewpoint of program theory. *ACM Trans. Inf. Syst.*, 33(1):3:1–3:30, March 2015. ISSN 1046-8188.

[17] J. Jiang, D. He, and J. Allan. Comparing In Situ and Multidimensional Relevance Judgments. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 405–414, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5022-8.

[18] Jiepu Jiang, Daqing He, and James Allan. Searching, Browsing, and Clicking in a Search Session: Changes in User Behavior by Task and over Time. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 607–616, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2257-7. doi: 10.1145/2600428.2609633. URL http://doi.acm.org/10.1145/2600428.2609633.

[19] K. Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[20] M. Kellar, C. Watters, J. Duffy, and M. Shepherd. Effect of task on time spent reading as an implicit measure of interest. In *Proc. 67th Annual Meeting Am. Soc. Inf. Sci. Technol.*, pages 168–175, 2004.

[21] D. Kelly. Implicit feedback: using behavior to infer relevance. *New directions in cognitive information retrieval*, pages 169–186, 2005.

[22] D. Kelly and N. J. Belkin. Display time as implicit feedback: understanding task effects. In *Proc. SIGIR'04*, pages 377–384. ACM, 2004.

[23] J. Kim, J. Teevan, and N. Crasswell. Explicit in situ user feedback for web search results. In *Proc. SIGIR'16*, pages 829–832. ACM, 2016.

[24] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In *Proc. WSDM'14*, pages 193–202. ACM, 2014.

[25] C. Liu, J. Gwizdka, and J. Liu. Helping identify when users find useful documents: examination of query reformulation intervals. In *Proceedings of IIiX'10*, pages 215–224, 2010.

[26] C. Liu, N.J. Belkin, and M.J. Cole. Personalization of search results using interaction behaviors in search sessions. In *Proc. SIGIR'12*, pages 205–214. ACM, 2012.

[27] J. Liu and N. J. Belkin. Personalizing information retrieval for multi-session tasks: the roles of task stage and task type. In *Proc. SIGIR'10*, pages 26–33. ACM, 2010.

[28] J. Liu and N.J. Belkin. Searching vs. writing: Factors affecting information use task performance. *Proc. Am. Soc. Inf. Sci. Technol.*, 49(1):1–10, 2012.

[29] Jingjing Liu and Nicholas J. Belkin. Personalizing information retrieval for multi-session tasks: Examining the roles of task stage, task type, and topic knowledge on the interpretation of dwell time as an indicator of document usefulness. *Journal of the Association for Information Science and Technology*, 66(1):58–81, January 2015. ISSN 2330-1643. doi: 10/gdj32k.

[30] Jingjing Liu, Chang Suk Kim, and Caitlin Creel. Exploring search task difficulty reasons in different task types and user knowledge groups. *Information Processing & Management*, 51(3):273–285, May 2015. ISSN 0306-4573. doi: 10/f66ssv.

[31] J. Mao, Y. Liu, K. Zhou, J. Nie, M. Zhang, and S. Ma. When does relevance mean usefulness and user satisfaction in web search. In *Proc. SIGIR'16*, pages 463–472. ACM, 2016.

[32] J. Mao, Y. Liu, H. Luan, M. Zhang, S. Ma, H. Luo, and Y. Zhang. Understanding and Predicting Usefulness Judgment in Web Search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1169–1172, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5022-8.

[33] Sophie Monchaux, Franck Amadieu, Aline Chevalier, and Claudette Mariné. Query strategies during information searching: Effects of prior domain knowledge and complexity of the information problems to be solved. *Information Processing & Management*, 51(5):557–569, September 2015. ISSN 0306-4573. doi: 10/f7npjg.

[34] D. Odijk, R. W. White, A. Hassan Awadallah, and S. T. Dumais. Struggling and Success in Web Search. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1551–1560, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3794-6.

[35] M. Potthast, M. Hagen, B. Stein, J. Graßegger, M. Michel, M. Tippmann, and C. Welsch. ChatNoir: A Search Engine for the ClueWeb09 Corpus. In Bill Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, *35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12)*, page 1004. ACM, August 2012. ISBN 978-1-4503-1472-5.

[36] M. Potthast, M. Hagen, M. Völske, and B. Stein. Crowdsourcing Interaction Logs to Understand Text Reuse from the Web. In Pascale Fung and Massimo Poesio, editors, *Proc. ACL'13*, pages 1212–1221. Association for Computational Linguistics, August 2013.

[37] Martin Potthast, Matthias Hagen, Michael Völske, Jakob Gomoll, and Benno Stein. Webis Text Reuse Corpus 2012, September 2012. URL https://doi.org/10.5281/zenodo.1341602.

[38] T. Sakai and Z. Dou. Summaries, ranked retrieval and sessions: A unified framework for information access evaluation. In *Proc. SIGIR'13*, pages 473–482. ACM, 2013.

[39] M.D. Smucker and C. Jethani. Time to judge relevance as an indicator of assessor error. In *Proc. SIGIR'12*, pages 1153–1154. ACM, 2012.

[40] A. Spink and T. Saracevic. Interaction in information retrieval: selection and effectiveness of search terms. *J. Am. Soc. Inf. Sci. Technol.*, (8):741–761, 1997.

[41] Lynda Tamine and Cecile Chouquet. On the impact of domain expertise on query formulation, relevance assessment and retrieval performance in clinical settings. *Information Processing & Management*, 53(2):332–350, March 2017. ISSN 0306-4573. doi: 10/f9qrkq.

[42] P. Vakkari. Task based information searching. *ARIST*, (37):413–464, 2003.

[43] P. Vakkari and S. Huuskonen. Search Effort Degrades Search Output But Improves Task Outcome. 63(4):657–670, 2012. ISSN 1532-2882.

[44] P. Vakkari, M. Völske, M. Potthast, M. Hagen, and B. Stein. Predicting retrieval success based on information use for writing tasks. In *22nd International Conference on Theory and Practice of Digital Libraries, TPDL 2018, Porto, Portugal, September 10-13, 2018, Proceedings*, 2018.

[45] B.M. Wildemuth, R. de Bliek, C.P. Friedman, and D.D. File. Medical students' personal knowledge, searching proficiency, and database use in problem solving. *Journal of the American society for Information Science*, 46 (8):590, 1995.

[46] E. Yilmaz, M. Verma, N. Craswell, F. Radlinski, and P. Bailey. Relevance and effort: an analysis of document utility. In *Proc. CIKM'14*, pages 91–100. ACM, 2014.

[47] Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. Predicting User Knowledge Gain in Informational Search Sessions. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 75–84, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5657-2. doi: 10.1145/3209978.3210064.