Received: 20 July 2020 | Revised: 16 November 2020 | Accepted: 20 December 2020

DOI: 10.1002/asi.24451

RESEARCH ARTICLE

JASIST WILEY

Predicting essay quality from search and writing behavior

Pertti Vakkari[1] | **Michael Völske**[2] | **Martin Potthast**[3] | **Matthias Hagen**[4] | **Benno Stein**[2]

[1]Tampere University, Tampere, Finland

[2]Bauhaus-Universität Weimar, Weimar, Germany

[3]Leipzig University, Leipzig, Germany

[4]Martin-Luther-Universität Halle-Wittenberg, Halle, Germany

Correspondence
Michael Völske, Bauhaus-Universität Weimar, Weimar 99423, Germany.
Email: michael.voelske@uni-weimar.de

Abstract

Few studies have investigated how search behavior affects complex writing tasks. We analyze a dataset of 150 long essays whose authors searched the ClueWeb09 corpus for source material, while all querying, clicking, and writing activity was meticulously recorded. We model the effect of search and writing behavior on essay quality using path analysis. Since the boil-down and build-up writing strategies identified in previous research have been found to affect search behavior, we model each writing strategy separately. Our analysis shows that the search process contributes significantly to essay quality through both direct and mediated effects, while the author's writing strategy moderates this relationship. Our models explain 25–35% of the variation in essay quality through rather simple search and writing process characteristics alone, a fact that has implications on how search engines could personalize result pages for writing tasks. Authors' writing strategies and associated searching patterns differ, producing differences in essay quality. In a nutshell: essay quality improves if search and writing strategies harmonize—build-up writers benefit from focused, in-depth querying, while boil-down writers fare better with a broader and shallower querying strategy.

1 | INTRODUCTION

Users of information systems seek support in various daily activities, and while search processes, that is, querying and assessment of search results, likely affect the outcome of the tasks behind the search, only few studies have explored these associations. The worth of the search to the user depends on their benefit in achieving a task outcome—be it a decision, a learning endeavor, or text being written. Search systems are typically evaluated based on the quality of their result lists, where quality refers to the number and position of relevant information items on the list. In other words, performance is assessed by the *output* of the search, rather than by its *outcome*, that is, the resulting benefits to the task at hand (Belkin, 2010; Järvelin et al., 2015; Vakkari, 2003). While search process, output, and outcome are generally assumed to be associated, the results from the few studies dealing with this problem are contradictory and partial; how and to what extent the information search process contributes to the quality of task outcomes is still an open question.

Since the utility of search results ultimately depends on how much they contributes to the outcome of an underlying task, the use of information retrieved is a link between search process and task outcome (Järvelin et al., 2015; Vakkari, Völske, Potthast, Hagen, & Stein, 2019). How found information is used reflects its utility, but only few studies have explicitly analyzed this (Vakkari, 2020). Our study takes information use into account and connects it both to the search process and to task outcome. We analyze to what extent search process—i.e., querying, result examination, and information selection in documents—and text

writing process are jointly associated to the quality of an essay text. Using path analysis, we reveal each factor's direct and mediated effects on essay quality. All told, our study has three unique contributions: (a) It shows how search and writing process are associated to task outcome by (b) analyzing their joint effects on task outcome, and (c) analyzes in this context the use of information found during search.

Specifically, our study examines the associations between querying, search result examination, and writing effort on the one hand, and the quality of an essay text on the other. We recruited 12 participants to write essays on 150 different topics and collected detailed logs of their searching and writing activity. We form querying, clicking, and writing effort constructs via factor analysis and build path models to predict the variation in essay quality. Previous studies have identified two distinct writing strategies in this context and shown that they have a significant impact on writers' querying and clicking behavior, and consequently, likely on the task outcome: build-up writers collect material aspect by aspect and construct the essay accordingly, while boil-down writers focus first on gathering a lot of material, which they then edit heavily (Hagen, Potthast, Völske, Gomoll, & Stein, 2016; Potthast, Hagen, Völske, & Stein, 2013).

Based on the hypothesis that different mechanisms of search, essay composition, and task outcome apply, we model the two writing strategies separately. Our results show that search process factors have significant direct and mediated effects and that writing factors have significant direct effects on essay quality for both writing strategies. Build-up and boil-down writing produce differences in essay quality; furthermore, writers who harmonize their search and writing strategies achieve higher-quality essays: while high-effort and low-diversity queries produce better essays for build-up writers, the opposite is the case for boil-down writers.

2 | RELATED WORK

Only few studies investigate how the information search process shapes task outcome. For instance, Hersh (1994) has been an early advocate for evaluating search systems from a task perspective: in addition to topical relevance, he proposed that systems be ultimately evaluated by their outcomes to users. Several subsequent studies on how IR systems can help students answer factual questions found no association between search output and task outcome: while participants answer questions better after searching than before, the size and quality of the result set had no significant impact (Hersh, 2003; Hersh, Pentecost, & Hickam, 1996). Another early study by Wildemuth, de Bliek, Friedman, and File (1995) analyzed to what extent search systems contribute to students' ability to answer clinical problems, and how search proficiency is associated with database-assisted problem solving. Here, both proficiency and search output were positively correlated with better responses.

A later study by Vakkari and Huuskonen (2012) on how medical students search for information for clinical-question essays measured the fraction of retrieved documents cited in the essay, along with the quality of the essay as assessed by experts. Effort in querying and exploring result documents was found to *degrade* precision (i.e., search output) but to improve task outcome. Outside the medical domain, Liu and Belkin (2012) studied the relationship between task type, search behavior, and task outcome in report writing and found that effort invested in writing did produce better reports. However, the query and click variables analyzed were not associated to outcome measures.

A study by Collins-Thompson, Rieh, Haynes, and Syed (2016) found that result exploration and time devoted to reading documents contribute more to learning outcomes than querying strategies. More recently, Yu et al. (2018) studied users' knowledge gain through search and concluded that dwell time in documents predicts users' knowledge gain, while their skill at querying and selecting results reveals more about their knowledge state.

As the above makes clear, findings on the relationship between search and task outcome are highly varied; precision and recall either have no effect (Hersh, 2003), have a positive effect (Wildemuth et al., 1995), or have a negative effect (Vakkari & Huuskonen, 2012) on task outcome. This may in part be due to the differences in tasks such as question answering or essay writing. Results on the contribution of search process variables to task outcome vary, as well: effort in the search process either improves task outcome (Vakkari & Huuskonen, 2012) or has no effect on it (Bron et al., 2012; Liu & Belkin, 2012). However, some studies indicate that the more effort is invested in reading (Collins-Thompson et al., 2016; Yu et al., 2018), or in writing compared to searching, the better the quality of a resulting essay (Liu & Belkin, 2012; Vakkari & Huuskonen, 2012).

Based on the same dataset as the one used in our own study, Vakkari et al. investigated the effect of search behavior on task outcome as measured by writers' success at retrieving sources for an essay and found search behavior characterized by clicking many varied results to be most indicative of success (Vakkari, Volske, Hagen, Potthast, & Stein, 2018). Dwell times had a negative effect on retrieval success, negatively mediated by clicking activity (Vakkari et al., 2019). The following investigation aims to

further deepen our understanding of the relationship between search and writing behavior on the one hand, and their impact on the task outcome on the other. Rather than using retrieval success as the outcome measure, we estimate the quality of the resulting essay.

3 | THE WEBIS-TRC-12 DATASET

Our analysis is based on the Webis Text Reuse Corpus 2012 (Potthast, Hagen, Völske, Gomoll, & Stein, 2012) which comprises 150 essays written by 12 writers; after briefly recapitulating the data acquisition process described by (Potthast et al., 2013), we detail the predictors we derive from these data for our own study.

3.1 | Data acquisition

The essays in the Webis-TRC-12 dataset were written in response to 150 different writing prompts, each in turn based on a topic from the TREC Web Tracks—Potthast et al. (2013) give an example. Participants were recruited from an online crowdsourcing platform, instructed to select a topic, and write an essay of approximately 5,000 words, while using a search engine to gather material. Text reuse from the sources found was encouraged. Participant demographics are reproduced in Table 1—the "typical" writer was a middle-aged, well-educated, native-English speaking woman with 8 years of professional writing experience.

The study was set up to record as detailed a picture of the search and writing process as possible: writers worked in a static web search environment consisting of a search engine indexing the ClueWeb09 crawl (Potthast et al., 2012), and a web emulator that, for a given address, returns the corresponding web page from the crawl. Furthermore, writers wrote their essays using an online rich text editor, which kept a fine-grained revision history, storing a new revision whenever the 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. The dataset's constructors had several research tasks in mind to which the corpus could contribute, including the study of search behavior in complex, exploratory information retrieval tasks (Hagen et al., 2016), the study of plagiarism and retrieving its sources (Hagen, Potthast, & Stein, 2015), and the study of writing behavior when reusing text and paraphrasing it (Potthast et al., 2013). Since the writers were asked to reuse parts of relevant web pages rather than write original text from scratch, the evolution of copied and pasted text passages can be tracked revision by revision, as can its contribution to the final essay, which was unobservable in most other studies. Each essay combines material from up to dozens of web pages, organized into a coherent final text.

TABLE 1 Demographics of the 12 writers (taken with permission from Potthast et al., 2013)

Writer demographics					
Age		**Gender**		**Native language(s)**	
Minimum	24	Female	67%	English	67%
Median	37	Male	33%	Filipino	25%
Maximum	65			Hindi	17%
Academic degree		Country of origin		Second language(s)	
Postgraduate	41%	United Kingdom	25%	English	33%
Undergraduate	25%	Philippines	25%	French	17%
None	17%	United States	17%	Afrikaans, Dutch	
n/a	17%	India	17%	German, Spanish	
		Australia	8%	Swedish each	8%
		South Africa	8%	None	8%
Years of writing		Usual search engines		Search frequency	
Minimum	2	Google	92%	Daily	83%
Median	8	Bing	33%	Weekly	8%
Standard deviation	6	Yahoo	25%	n/a	8%
Maximum	20	Others	8%		

3.2 | Variables

In order to model the relationship between writer behavior and essay quality, we compute 26 numerical variables for each of the 150 essays, distributed among four categories: (a) querying behavior, (b) click and reading behavior, (c) writing effort, and (d) essay quality of the finished text. Table 2 gives basic statistics; unless otherwise indicated below, each variable is measured as an absolute frequency, aggregated over the entire time a writer spent working on a single essay. The table columns characterize the distribution over the 150 essays.

The first row group shows query variables: in order to produce a typical essay, writers submitted 34.5 queries, 20 of them unique. *Anchor queries* refer to queries that are resubmitted occasionally, as a way for a writer to keep track of the main theme of their task (Hagen et al., 2016). We subdivide a writer's work on the task into physical sessions, whenever a break of 30 min or more occurs; *search sessions* refer to sessions during which the writer submitted at least one query. For the terms that make up a writer's queries, we count both their total number and the number of unique terms. To quantify how writers adapt their queries while working, we record the number of unique query terms that were

TABLE 2 Independent variables on queries, clicks, and writing (top three row groups), and dependent variables on essay quality (bottom row group) per essay ($n = 150$)

Variable	Min	Med	Max	μ	σ
Querying behavior (independent)					
Queries	2.0	34.5	307.0	46.9	42.2
Unique queries	1.0	20.0	122.0	24.5	17.6
Anchor queries	0.0	3.0	33.0	5.7	6.5
Search sessions	1.0	6.5	24.0	7.1	4.0
Sessions	1.0	10.0	34.0	11.0	6.7
Query terms	4.0	193.0	1946.0	260.3	258.8
Unique terms (UT)	1.0	24.0	195.0	29.8	25.1
UT from snippets	1.0	21.0	188.0	26.4	23.3
UT from documents	1.0	20.0	179.0	24.8	21.9
Querying time (min)	0.5	24.2	266.5	40.8	45.5
Click and reading behavior (independent)					
Clicks	12.0	88.0	443.0	112.9	81.1
Useful clicks	0.0	25.5	122.0	32.5	25.7
Pastes	0.0	25.0	134.0	28.0	21.4
Words pasted	306.0	7376.5	33542.0	8709.9	5371.6
Click trails	5.0	47.5	280.0	59.3	42.2
Avg. click trail depth	0.0	1.6	20.4	2.0	2.0
Reading time (min)	11.8	60.0	380.1	81.3	66.1
Writing effort (independent)					
Revisions	249.0	2830.5	6948.0	2826.8	1422.9
Major revisions	6.0	32.0	92.0	33.2	16.1
Final word count	717.0	4922.5	14088.0	4987.2	1283.1
Final sources count	0.0	16.0	69.0	17.7	12.4
Writing time (min)	38.5	317.2	1379.5	363.6	226.3
Essay quality (dependent)					
Organization	2.4	3.6	4.7	3.6	0.3
Thesis clarity	2.9	3.7	4.4	3.7	0.3
Prompt adherence	2.4	3.4	4.5	3.4	0.4
Argument strength	2.8	4.1	5.1	4.1	0.4

Note: The embedded plots visualize variables' frequency distributions, scaled between the respective min and max (frequency) on the *x*-axis (*y*-axis).

first submitted after they occurred in a SERP snippet or a document that the writer saw.

The second row group shows click variables. Next to the total number of clicks (on search result page entries, or on links inside search results), we also record the number of *useful* clicks. These are clicks on documents from which at least one passage of text is pasted into the essay. We further record the number of times such *pastes* took place, as well as the combined number of words pasted from search results—the large variance is notable here, but Potthast et al. (2013) have already discussed this phenomenon. A *click trail* refers to a sequence of documents visited after selecting a search result, and then following hyperlinks from one document to the next; we record both the *number* of click trails and their average depth. Finally, we record the combined time spent reading the retrieved documents.

The third row group shows writing effort variables, including the number of revisions recorded during essay writing (recall that a new revision is recorded whenever the writer stops typing for more than 300 ms), as well as the number of *major* revisions that alter at least 5% of the text. As also noted in previous work, essay length may fluctuate considerably over time. For the purposes of our study, we consider only the final word count.

The bottom row group of Table 2 shows four essay quality dimensions proposed by Persing, Davis, and Ng (2010) and Persing and Ng (2013, 2014, 2015) for which Wachsmuth, Khatib, and Stein (2016) show that they can be measured using argument mining at state-of-the-art accuracy; we apply the latter's approach to the Webis-TRC-12 essays. The *organization* score rates the structure of an essay. A well-organized essay introduces a topic, states a position, supports the position, and concludes—it logically develops an argument (Persing et al., 2010). *Thesis clarity* evaluates how clearly an essay explains its overall message or position; a high-scoring essay presents its thesis in an easily-understandable way (Persing & Ng, 2013). The *prompt adherence* score is high for essays that consistently remain on the topic of the writing prompt (Persing & Ng, 2014). Finally, an essay scores high along the *argument strength* dimension if it makes a convincing argument for its thesis (Persing & Ng, 2015). We compute the final essay quality scores by standardizing each of these dimensions as z scores and then adding up the results.

4 | MODELING ESSAY QUALITY

We analyze the connections of query, click, and writing effort variables to essay quality, forming path models to measure direct and mediated effects, as well as group differences based on writing strategy. Our statistical model identifies the effect of each independent variable on the variance of a dependent variable and thus indicates the relative effect of each variable on other variables.

4.1 | Path analysis

We use path analysis instead of structural equation modeling due to the small number of cases in our data. The latter requires hundreds of cases in order to be reliable. Path analysis—a special case of structural equation modeling—is a technique for describing the directed dependencies between variables in a regression model. In addition to the direct effects of independent on dependent variables, path analysis also models the effects of each independent variable on the others, and thus the indirect effects on the dependent variables. Relations among variables are expected to be linear and without interactions Hair, Black, Babin, & Anderson, 2010). The direct effects between pairs of variables are characterized by path coefficients (β): standardized regression coefficients obtained through a series of ordinary regression analyses, where each variable is taken as the dependent variable in turn.

Modeling essay quality by query, click, and writing effort variables requires an understanding of the process, that is, the causal order of search and writing variables and how they may be interrelated. The literature gives some hints to the nature of these dependence relationships (e.g., Järvelin et al., 2015; Vakkari, 2020). Järvelin et al. (2015) have described in detail how task planning, searching information items, selecting between them, working with them, and synthesizing and reporting are associated with task outcome. They hypothesize based on the literature how these five generic activities contribute to task performance and outcome. Hagen et al. (2016) have analyzed (using also the Webis-TRC-12 data) how writers search and how search process and outcome may be associated. Based on the literature, Vakkari (2020) has systematized how the search process is associated to the usefulness of search results. Together, these three studies provide sufficient information about the associations between search and task performance process to inform our path models.

Since a multitude of search variables may affect essay quality, we use constructs that combine several individual variables under a concept seeking to cover the variety of these variables. Using constructs also simplifies the analysis by reducing the number of predictors in the models (Hair et al., 2010). Since our goal is to compare the strategies build-up and boil-down writers use to

achieve a high-quality essay, we model writers with each writing strategy separately.

4.2 | Forming subgroups by writing strategy

Previous work that used the Webis Text Reuse Corpus 2012 to study the search and writing strategies of essay writers has found evidence of two distinct writing strategies—boil-down and build-up—characterized by broad, up-front versus selective, on-demand material gathering (Hagen et al., 2016; Potthast et al., 2013). Build-up writers work aspect by aspect and grow the essay continuously, alternating short, targeted material gathering and re-writing sessions. Boil-down writers collect lots of material in big bursts early in the search process and then switch to a re-writing phase characterized by little searching or new material gathering; consequently, their essays grow quickly at first and then contract again as the material is distilled.

Based on the aforementioned properties, Potthast et al. (2013) and Hagen et al. (2016) categorize each of the 150 essays into build-up, boil-down, or, if neither strategy clearly dominates, mixed. This categorization is carried out manually, by inspecting the development of the essay length, the distribution of pastes over time, and the points at which new sources are introduced. We adopt the categorization from the aforementioned previous works. Table 3 shows the number of essays in the dataset across authors and writing strategies: of the 12 authors, 5 use build-up, 2 use boil-down, and 5 use the mixed strategy most frequently. Authors do not adhere to the same writing strategy all the time but tend to show a rather clear preference: 8 of the 12 authors—who wrote two thirds of all essays—use their "favored" strategy for 70% or more of their essays. Out of the 10 authors who wrote more than one essay, only 3 favor the mixed strategy.

An initial examination reveals that the build-up and boil-down writing strategies produce significant models for essay quality, but the mixed writing strategy does not. Therefore, we remove essays with a mixed strategy ($n = 42$) from our analysis. Excluding the mixed essays, seven remaining authors predominantly use the build-up strategy, and three the boil-down strategy; one author uses both strategies equally.

4.3 | Construct development and factor analysis

While our raw data yield a large number of individual measurements of the authors' querying, result examination, and writing behavior (cf. Table 2), we aim to limit the number of variables included in our models for the sake of interpretability. To this end, we combine multiple measurements into constructs by way of factor analysis. Many concepts relevant to the study of information retrieval processes are in fact theoretical constructs like query quality or result list quality, which cannot be directly observed, but can be measured indirectly using multiple variables through factor analysis, where each factor represents a construct.

TABLE 3 Number of essays by writing strategy by author

	Essays		Writing strategy			
Author	Total (retained)		Build-up	Mixed	Boil-down	% maj.
u006	7	(7)	7	0	0	100.0
u014	1	(1)	0	0	1	100.0
u025	1	(0)	0	1	0	100.0
u020	10	(9)	9	1	0	90.0
u005	18	(15)	15	3	0	83.3
u002	33	(27)	1	6	26	78.8
u021	12	(10)	9	2	1	75.0
u018	20	(15)	14	5	1	70.0
u007	12	(5)	4	7	1	58.3
u024	11	(5)	1	6	4	54.5
u001	2	(2)	1	0	1	50.0
u017	23	(12)	10	11	2	47.8

Note: The final column shows the percentage of essays in which the author used their dominant writing strategy.

We select indicators for the major stages of the search and writing process (querying, result examination, and writing effort) and run several factor analyses with the aim of finding two or three factors per stage, based on the following three criteria: factors should use few variables with high communality, the variables in a factor should have high loadings, and the conceptual meaning of each factor should be clear. We use principal component analysis with varimax rotation, applied jointly to both writer groups ($n = 108$), and extract factors with an eigenvalue of at least 1.0, yielding the factors shown in Table 4.

Factors 1.1 and 1.2 represent query effort and query diversity, explaining 87.7% of the total variance of their constituent variables. Factor 1.1 indicates the amount of effort put into querying, both in terms of time investments and of its consequences, while Factor 1.2 combines variables that reflect the diversity and uniqueness of queries. The reliability of variables in both factors is high, varying from 0.76 to 0.85 in 1.1, and from 0.74 to 0.92 in 1.2. Factors 2.1, 2.2, and 2.3 together cover clicking and result examination, explaining 84.5% of the total variance of their constituent variables. Factor 2.1 reflects click utility—to what extent clicks are useful in terms of providing material for writing. Factor 2.2 is called paste volume, representing the amount of text extracted for further use. Factor 2.3 reflects the effort invested in examining results per query. The reliability of the constituent variables of these factors is high, varying between 0.77 and 0.88 with the exception of the number of pastes in Factor 2.2, which reaches only 0.50, but which we consider still high enough to include this variable in the construct. Factors 3.1 and 3.2 cover writing effort, explaining 83.3% of the total variance of their constituent variables. Factor 3.1 reflects the volume of revisions in writing the essay, while Factor 3.2 indicates writing effort per paste. The reliability of the variables in both constructs is high, varying from 0.72 to 0.90.

TABLE 4 Factor analysis for constructs (loadings >.40, $n = 108$)

Factors and constituent variables	Loading (L)	Reliability (L^2)
(1.1) Query effort		
Number of queries	0.92	0.85
Number of query terms (log)	0.91	0.83
Seconds spent querying	0.87	0.76
(1.2) Query diversity		
Unique terms per query	0.96	0.92
Unique terms from results per query	0.94	0.88
Percent unique queries	0.86	0.74
(2.1) Click utility		
Pastes per click (log)	0.89	0.79
Percent useful clicks	0.88	0.77
(2.2) Paste volume		
Words pasted	0.91	0.83
Number of pastes	0.70	0.50
(2.3) Click effort per query		
Clicks per query	0.94	0.88
Reading time per query (log)	0.93	0.86
(3.1) Revision volume		
Number of revisions	0.89	0.79
Writing time	0.85	0.72
(3.2) Writing effort per paste		
Writing time per paste (log)	0.95	0.90
Revisions per paste (log)	0.89	0.79

We calculate factor scores for each essay and average them across the build-up and boil-down groups. As shown in Table 5, the querying, clicking, and writing behaviors of both groups differ significantly. In build-up essays, the writers' query effort is significantly smaller, while query diversity is significantly greater compared to boil-down essays. There is no significant difference in click effort per query, but click utility is significantly greater and paste volume significantly smaller for build-up writers. Both groups invest similar amounts of writing effort per paste, while revision volume is significantly greater with the boil-down strategy. In all, it seems that the boil-down strategy requires more effort in querying, more pasting, and more revisions, while the build-up strategy achieves greater query diversity and click utility.

5 | RESULTS

Following the scheme outlined in the preceding section, we build separate path models for build-up and boil-down essays. We present only associations with $p < .10$ in the figures. The following three sections describe the models and highlight commonalities and differences of interest.

5.1 | Build-up essays

The path model for build-up essays is shown in Figure 1: independent variables are numbered as in Tables 4 and 5. Significant paths are shown with their β-coefficients and annotated with significance levels.[1] Positive correlation coefficients are highlighted in green, negative ones in red. The model is significant ($R^2 = .32$, Adj$R^2 = .25$,

TABLE 5 Means of factor scores by writing strategy

Factor	Build-up ($n = 71$)	Boil-down ($n = 37$)	p (t test)	t	df
(1.1) query effort	−0.218	0.418	0.006	2.87	51.95
(1.2) query diversity	0.259	−0.497	0.000	3.76	62.75
(2.1) click utility	0.213	−0.385	0.001	3.27	91.46
(2.2) paste volume	−0.321	0.581	0.000	4.73	68.48
(2.3) click effort per query	0.088	−0.159	0.195	1.31	91.95
(3.1) revision volume	−0.268	0.492	0.001	3.47	51.21
(3.2) writing effort per paste	0.044	−0.081	0.467	0.73	102.79

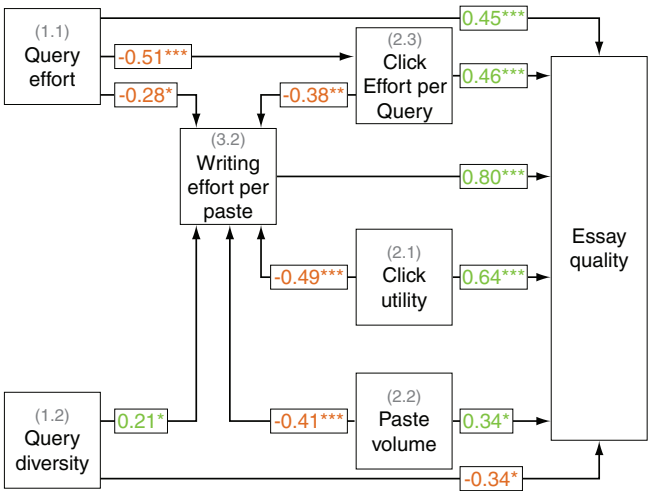

FIGURE 1 A path model for build-up essay quality ($n = 71$) [Color figure can be viewed at wileyonlinelibrary.com]

$F = 4.70$, $p = .001$) and explains 25% of the variation in the essay quality. All factors in the model have significant direct effects on essay quality.

While the three click factors have positive direct effects, their effects when mediated by writing effort per paste are negative: thorough inspection of search results and copious pasting directly improve essay quality, but this behavior reduces writing effort, and through that path decreases essay quality. Copious clicking and pasting seems to be directly associated with increased essay quality, while scant clicking and pasting improves essay quality after editing each paste with effort.

It seems as if frequent pasting produces a good essay without association to editing, while scant pasting implies the author has to work hard with the pasted text. Considering the proportion of original words—those not copy-pasted from sources—in the essay, all three click factors have a negative effect: click effort per query ($r = -.38$, $p < .001$), click utility ($r = -.40$, $p = .001$), and paste volume ($r = -.49$, $p < .001$) reduce the proportion of original words. By contrast, writing effort per paste increases the proportion of original words ($r = .59$, $p < .001$). Thus, some writers appear to produce good essays through abundant pasting without much editing, while others utilize search results more selectively and invest editing effort to improve their essays.

Interestingly, increasing query effort both directly and indirectly contributes to essay quality: on the one hand, increasing query effort decreases click effort per query, which in turn increases writing effort per paste, and through this path improves essay quality. Simultaneously, an increase in query effort also increases essay quality directly. This may imply that those who put effort into querying formulate pertinent queries, which produce good result lists that do not require much effort—in terms of clicks or reading time—to find useful documents. We can put this line of reasoning to the test since, after finalizing their essays, writers rated the quality of the top search results on a four-point scale from very useful to spam. A positive association between query effort and the rating of the top results would support the hypothesis that effort invested querying saves effort down the line. However, the correlation is negative ($r = -.35$, $p = .005$); thus, it seems more likely that increasing query effort reflects difficulties in formulating effective queries, whereas useful search results still require effort in result examination.

Writers also rated on a three-point scale how difficult it was to find useful sources, and this difficulty was significantly correlated with query effort ($r = .44$, $p < .001$). Like the finding above, this supports the hypothesis that query effort seems to reflect difficulty in query formulation, and consequently in finding useful sources for the essay. By way of further evidence, query effort is negatively associated with the number of pastes per query ($r = -.45$, $p < .001$) and with the number of words pasted per query ($r = -.57$, $p < .001$), but it correlates positively with reading time per paste ($r = .31$, $p = .01$). Thus, those who invest effort in querying spend lots of

time reading to select text to copy-paste, but select relatively few pastes and words pasted per query—likely due to difficulties in formulating queries that produce useful search results.

The interpretation that query effort reflects struggling to formulate effective queries also helps interpret the paths that link query effort to essay quality: On the first path, increasing query effort directly improves essay quality and indirectly reduces click effort, which increases writing effort, thus improving essay quality. On the second path, decreasing query effort increases both click and writing effort, both increasing essay quality. Thus, it seems that those who struggle in querying find only a limited amount of useful material; this causes extra editing effort to produce a good essay. By contrast, some writers find useful search results with less query effort, and do not require much editing, but still produce a good essay.

5.2 | Boil-down essays

The path model for boil-down essays (Figure 2) is significant ($R^2 = .41$, Adj$R^2 = .35$, $F = 7.00$, $p = .001$) covering 35% of the variation in essay quality. One factor in the model, query effort, contributes significantly to essay quality, while click effort per query and revision volume have a notable effect on essay quality. There is only one mediated effect on essay quality: a decrease in query effort increases click effort per query, which in turn increases essay quality. Simultaneously, a decrease in query effort also directly increases essay quality. Thus, it seems that decreasing query effort leads to increasing click effort, which then enhances essay quality.

For boil-down essays, time spent formulating queries hurts essay quality. A possible explanation is that as in build-up essays, query effort reflects problems in formulating effective queries, leading to additional effort in finding useful material for the essay in poor result lists, and finally to lower scores in essay quality. We check the validity of this hypothesis, once again using the writers' subjective ratings of search result quality: the rating of the top search results was negatively, but not significantly

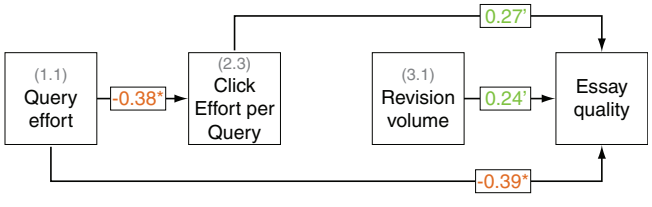

FIGURE 2 A path model for boil-down essay quality ($n = 37$) [Color figure can be viewed at wileyonlinelibrary.com]

associated with query effort ($r = -.20$, $p = .25$). Although the direction of association supports the hypothesis, there is insufficient evidence that query effort reflects difficulties in finding useful search results.

To supplement the above analysis, we consider the fraction of the number of paste clicks out of all clicks as an additional indicator of good result lists as perceived by writers. The higher this ratio, the better the result lists. If the association between query effort and the fraction of paste clicks is negative, the hypothesis that querying effort reflects struggling gets support. The correlation is indeed negative ($r = -.48$, $p = .002$). The more effort in querying, the lower the proportion of paste clicks of all clicks. Moreover, the writers' perceived difficulty of finding useful sources significantly increases with an increase in query effort ($r = .32$, $p = .05$), which corroborates our hypothesis.

5.3 | Comparison of models

The models are structurally quite different, with the much simpler model for boil-down essays including fewer query and click factors. The model for boil-down essays covers 35% of the variation in essay quality, while the model for build-up essays covers only 25%.

Among the differences between the models, the opposite effect of the *query effort* factor on essay quality is particularly interesting: an increase in query effort directly decreases essay quality in boil-down essays, while it enhances essay quality in build-up essays. As noted previously, writers who struggle with querying find less useful search results and therefore tend to receive lower essay scores. The writers of build-up essays seem to directly compensate for this with more querying effort, whereas for boil-down writers, increasing query effort has a detrimental effect.

In build-up writers, effort in editing the essay text improves essay scores in conjunction with the effort in querying for finding useful sources: writing effort mediates between essay quality and query and click factors. Thus, it seems that effort in revising and editing the essay text compensates for the difficulties in finding useful search results, and in effect produces higher essay scores. Conversely, effort in revising boil-down essays does not mediate between essay quality and query and click factors—instead, it enhances essay quality regardless of query and click factors. In build-up essays, not only query effort but also query diversity is associated with essay quality, with an opposite effect: an increase in query diversity directly decreases essay scores.

On the whole, different querying strategies seem to lead to different outcomes for the two groups of writers.

TABLE 6 Mean essay quality scores by writing and querying strategy

	Query effort (1.1)		Query diversity (1.2)	
Writing strategy	**Low (≤0.7)**	**High (>0.7)**	**Low (≤ − 0.5)**	**High (> − 0.5)**
Build-up	0.1	1.1	1.3	0.1
	$n = 62$	$n = 9$	$n = 12$	$n = 59$
Boil-down	0.8	−1.5	−0.6	1.3
	$n = 23$	$n = 14$	$n = 26$	$n = 11$
p (t test)	0.18	0.01	0.02	0.13
t	1.36	2.82	2.49	1.61
df	59.07	12.99	37.00	15.70

For a deeper look into the interplay between writing strategy, querying strategy, and essay scores, we divide the build-up and boil-down essays along the two query strategy factors—query effort and query diversity—into subgroups with low and high factor scores and study the intergroup differences in mean essay scores. We analyze the association of the resulting categorization scheme with essay scores via two-way ANOVAs. None of the main effects of query effort ($F = 1.2$, $p = .28$), query diversity ($F = 0.3$, $p = .57$), and build-up ($F = 2.5$, $p = .12$), or boil-down ($F = 0.3$, $p = .59$) writing style are significant. However, there is a significant interaction effect between writing style and both query effort ($F = 7.5$, $p = .007$), and query diversity ($F = 7.2$, $p = .009$). Thus, depending on the writing strategy, different query strategies have different effects on essay quality.

Table 6 elaborates on this interaction and highlights the magnitude of the effect: high query effort produces significantly higher essay scores in the build-up writing strategy compared to the boil-down strategy, while low query effort produces better scores in boil-down compared to build-up writers. For query diversity, the mechanism is reversed: low query diversity is significantly associated with high essay scores in the build-up writing strategy, but with low scores in the boil-down strategy, whereas high query diversity produces low essay scores in the build-up group, but high scores in the boil-down group. Although the differences in essay scores between the writing strategies are not significant in the low query effort and high query diversity groups, altogether the differences are systematic, which speaks for their validity.

In sum, writers following the build-up strategy achieve better essay quality scores through high query effort and low-diversity queries, whereas boil-down writers fare better with an opposite querying strategy involving lower effort and higher-diversity queries. Given the way the two querying strategy factors are composed, this might imply that build-up writers profit more from exhaustively covering a narrower subset of the topic space with their queries, while boil-down writers do better with less in-depth queries, but covering a broader range of subtopics.

6 | DISCUSSION

Our study is one of the first attempts at modeling task outcome by search process variables taking into account the use of information found for the task. Its results significantly extend our understanding of how search process and writing effort contribute to the quality of essay text and how writing strategies moderate the contribution of these variables. Our findings have implications on retrieval system design, such as in search personalization or query suggestion.

6.1 | Search process affects task outcome

Our models show that the search process significantly affects the outcome of the underlying task: especially in build-up essays, query and click factors contribute both directly and indirectly to essay quality. While it is not obvious how queries, in particular, should directly affect essay quality, our models reveal that they do so via mediating factors like search result examination or text editing behavior. It seems plausible that query patterns reflect some latent factor—such as prior conception of the topic—influencing essay quality directly.

Search result examination factors also have a direct effect on essay quality: similarly to Vakkari and Huuskonen (2012), the effort invested in examining opened documents improves our writers' conception of the essay topic and thus essay quality. The effect reported by Collins-Thompson et al. (2016)—that more time spent reading documents improves participants' task outcomes—is also evident in our models. Additionally, for both writing strategies, writing and revision effort invested in the essay itself naturally have a direct positive contribution to essay quality.

Beyond the direct effects, query and click factors also affect essay quality indirectly, although the clarity and complexity of these mediated effects vary. For both writing strategies, decreasing query effort seems to indicate the writers' ability to formulate effective queries that directly contribute to essay quality. In both groups, increasing query effort leads to reduced click effort and thus to lower essay quality. As previously reported by Smith and Kantor (2008) and Awadallah, White, Dumais, and Wang (2014), query effort may indicate difficulty formulating effective queries here. However, writers following the build-up strategy can overcome the negative impact of increasing query effort, decreasing click effort, and decreasing pastes through increased writing effort: faced with a shortage of successful queries yielding useful sources and text fragments, these writers can still compensate by working harder at integrating the sources they do find. This corresponds to the findings by Liu and Belkin (2012): investing effort in writing instead of searching improves writing outcome.

However, one may suspect that due to the considerable variation in essay length and time spent writing essays, these variables may act as mediating factors. By implication, path models might vary between, for example, long and short essays. To test this suspicion, we calculate correlations between essay quality and word count ($r = -.13$, $p = .19$) and between essay quality and writing time ($r = -.01$, $p = .96$). Neither of the two factors is significantly associated with essay quality.

6.2 | Writing strategy moderates the relationship

One of the most interesting insights from our study relates to the interplay between querying tactics and writing strategy in producing high-quality essays: The querying strategy that contributes to high-quality essays for build-up writers negatively affects the quality of boil-down writers' essays, and vice versa. Build-up writers benefit from high query effort combined with low query diversity, whereas in boil-down essays a low query effort in conjunction with a strong query diversification led to high essay scores.

Build-up writers compile their essays gradually, searching information aspect by aspect; search and writing periods alternate until the essay is finished. This approach seems to require a preplanned structure for the essay. Boil-down writers search and collect information in bursts and tend to gather much more content than is needed for the essay text; an editing and shortening phase follows the search phase. It is likely that the structure of the essay is shaped while examining search results and while writing the essay. These two writing strategies resemble the Analytic and Wholist cognitive styles. The former sees a situation as a collection of parts and focus on few aspects at a time like build-up writers, while the latter tend to see a situation as a whole picture and retain an overall view of the information like the boil-down writers (Goodale, Clough, Fernando, Ford, & Stevenson, 2014; Kinley, Tjondronegoro, Partridge, & Edwards, 2014).

Our results indicate that the search strategies for compiling high-quality essays differ between the observed writing strategies (Table 6). Build-up writers invest effort in querying while keeping query diversity low. Boil-down writes invest in query diversity while keeping query effort low. Here again the build-up writing style is Analytic in the sense that it focuses on one aspect of the topic at a time (low query diversity) but requires more effort to formulate effective queries. The boil-down strategy is Wholistic and produces broad and diverse queries, seemingly without much effort. Thus, writing strategies likely reflect the Analytic–Wholistic dimension, along which different query tactics produce high-quality essays.

It may thus be a fruitful direction for retrieval systems to detect when users are engaged in writing tasks, and further, to detect which writing strategy users are pursuing, as it would seem prudent to support and guide different types of writers differently based on such knowledge. Given the prevalence of web-based writing tools, systems could learn to identify users' writing strategy automatically and adapt the search frontend accordingly. For instance, users working in a build-up style could receive long and specific query suggestions, whereas boil-down writers might see more topically diverse suggestions. Similarly, the scoring function used for ranking could be modified to prioritize relevance or diversity. Any search facets could be tailored to focus either on subtopics of the current query or on adjacent topics, instead.

Table 6 indicates that providing boil-down writers with diverse query suggestions representing new aspects of the topic would likely lead to better essay scores. During their initial material-gathering phase, boil-down writers stand to benefit especially from suggestions that—in the terminology of Raman, Bennett, and Collins-Thompson (2013)—cater to the intrinsic diversity of the task. By contrast, Hagen et al. (2016) found that build-up writers tend to compile their essays in a more targeted manner, going aspect by aspect. Query diversity by itself tends to be weakly detrimental—or at least not beneficial—to these writers' essays. However, when mediated by a greater writing effort per paste (see Figure 1), diverse queries can help even build-up writers.

Beyond the interaction with querying tactics, writing strategies also moderate the effect of the examination and

use of search results on essay quality. While in boil-down essays only search result examination (i.e., click effort) has a mediated effect on essay quality, build-up essays exhibit richer interactions, in which both search result examination and the amount of information extracted from the sources have direct and mediated effects on essay quality. Increases in both result examination and information extraction directly improve essay quality, but on the other hand, decreases in result examination and information extraction, mediated by increasing writing effort, can also improve essay scores. It seems that when examining search results, some writers are able to identify and extract lots of text fragments from sources that match the evolving text well, thus improving their essays. Other writers likely compensate for a scarcity of useful text passages by working harder on editing them to match the essay text.

Our models connect query, click, and writing effort factors to essay quality by various paths, suggesting that real-world search and writing processes are complex. A writer may invest effort in some phase and thereby influence the strength of associations elsewhere in the process (Smith & Kantor, 2008). Differences between writing strategies in searching may have effects that nullify each other, implying spurious interdependencies. Simulation models of search behavior seldom account for such effects and may thus include various sources of error.

6.3 | Limitations

While our study provides essential insights into how writers search and use information for a writing task, a series of important caveats must be noted: First, the number of participants is small, and the distribution of writing styles across writers is quite uneven; 26 out of the 37 boil-down essays (70%) were written by the same person. A t test indicates significant differences between essays by this writer (code-named "u002") and others in query diversity ($p = .001$), query effort ($p = .002$), and in click utility ($p = .009$), but not in click effort, paste volume, writing effort per paste, or revision volume. Follow-up studies will need to validate our findings at larger sample sizes.

Second, we average search and writing process variables over the entire writing process, which likely reduces the variation of the phenomenon under investigation, and in turn may decrease the value of association measures like correlations (Hair et al., 2010). A next step toward higher validity would use the search session as the unit of observation, which may reveal more clearly how patterned search tactics are associated with writing effort and essay quality.

Third, we use a limited set of automatic measures for essay quality; while our measures have been shown to be valid (Wachsmuth et al., 2016), more robust quality measures may be necessary. We envision studies using varying essay quality measures to confirm our results; supplementing the automatic measures with reviews by human experts may be especially desirable.

Fourth, the writers of our essays were encouraged to reuse text to complete their task, in contrast to many real-world settings where original writing is a strict requirement. However, we argue that writers will always seek out the most useful information possible, irrespective of originality requirements. Thus, querying and result examination would be more or less the same. Stricter originality requirements would affect the way writers select and edit text passages: in models like ours, the weight of writing effort would be greater, and the weight of the number of pastes smaller. Nevertheless, the models we have developed are plausible to a great extent, even in settings which do not allow text reuse in writing longer essays. In addition, text reuse does have a clear methodical benefit: writers' copying and pasting text passages makes their information use directly observable, in a more natural and less intrusive manner than alternative techniques could achieve—such as having participants rate the usefulness of individual search results.

7 | CONCLUSIONS

Our results show that query, click, and text editing factors have both direct and mediated effects on task outcome, in our case essay quality; in our models, these factors explain 25–35% of the variation in quality. Thus, the search process contributes significantly to the task outcome, and this contribution varies by writing strategy. Build-up and boil-down writers seem to benefit from different search strategies in pursuit of a high-quality essay, and future information retrieval systems might exploit this finding to better support writing tasks.

This potential notwithstanding, the varying paths to a high-quality essay in our models hint at the complexity of the dependencies between the search process and the task outcome. More theoretical and empirical work is required to elaborate and model these dependencies.

ACKNOWLEDGMENTS

Open Access funding enabled and organized by ProjektDEAL. WOA Institution: BAUHAUS-UNIVERSITAET WEIMAR. Blended DEAL: ProjektDEAL.

ORCID

Pertti Vakkari https://orcid.org/0000-0002-4441-5393
Michael Völske https://orcid.org/0000-0002-9283-6846
Martin Potthast https://orcid.org/0000-0003-2451-0665
Matthias Hagen https://orcid.org/0000-0002-9733-2890
Benno Stein https://orcid.org/0000-0001-9033-2217

ENDNOTE

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

REFERENCES

Awadallah, A. H., White, R. W., Dumais, S. T., & Wang, Y. (2014). Struggling or exploring?: Disambiguating long search sessions. In *7th annual international ACM conference on web search and data mining (WSDM 2014)* (pp. 53–62). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2556195.2556221

Belkin, N. J. (2010). On the evaluation of interactive information retrieval systems. In *The Janus faced scholar. A festschrift in honor of Peter Ingwersen* (Vol. 06-S, pp. 13–22). Copenhagen, Denmark: ISSI.

Bron, M., van Gorp, J., Nack, F., de Rijke, M., Vishneuski, A., & de Leeuw, S. (2012). A subjunctive exploratory search interface to support media studies researchers. In *The 35th international ACM SIGIR conference on research and development in information retrieval* (pp. 425–434). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2348283.2348342

Collins-Thompson, K., Rieh, S. Y., Haynes, C. C., & Syed, R. (2016). Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *CHIIR '16: Proceedings of the 2016 ACM on conference on human information interaction and retrieval* (pp. 163–172). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2854946.2854972

Goodale, P., Clough, P. D., Fernando, S., Ford, N., & Stevenson, M. (2014). Cognitive styles within an exploratory search system for digital libraries. *Journal of Documentation*, *70*(6), 970–996. https://doi.org/10.1108/JD-03-2014-0045

Hagen, M., Potthast, M., & Stein, B. (2015). Source retrieval for plagiarism detection from large web corpora: Recent approaches. In *Working Notes of CLEF 2015*. Conference and Labs of the Evaluation forum, Toulouse, France: CEUR Workshop Proceedings. Retrieved from http://ceur-ws.org/Vol-1391/inv-pap10-CR.pdf

Hagen, M., Potthast, M., Völske, M., Gomoll, J., & Stein, B. (2016). How writers search: Analyzing the search and writing logs of non-fictional essays. In *Proceedings of the 2016 ACM on conference on human information interaction and retrieval* (pp. 193–202). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2854946.2854969

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. (2010). *Multivariate data analysis*. Upper Saddle River, NJ: Prentice-Hall.

Hersh, W. R. (1994). Relevance and retrieval evaluation: Perspectives from medicine. *Journal of the American Society for Information Science*, *45*(3), 201–206.

Hersh, W. R. (2003). *Information retrieval: A health and biomedical perspective* (2nd ed.). New York, NY: Springer.

Hersh, W. R., Pentecost, J., & Hickam, D. H. (1996). A task-oriented approach to information retrieval evaluation. *Journal of the American Society for Information Science*, *47*(1), 50–56.

Järvelin, K., Vakkari, P., Arvola, P., Baskaya, F., Järvelin, A., Kekäläinen, J., ... Sormunen, E. (2015). Task-based information interaction evaluation: The viewpoint of program theory. *ACM Transactions on Information Systems*, *33*(1), 1–30. https://doi.org/10.1145/2699660

Kinley, K., Tjondronegoro, D., Partridge, H., & Edwards, S. L. (2014). Modeling users' web search behavior and their cognitive styles. *Journal of the Association for Information Science and Technology*, *65*(6), 1107–1123. https://doi.org/10.1002/asi.23053

Liu, J., & Belkin, N. J. (2012). Searching vs. writing: Factors affecting information use task performance. In *Proceedings of the American Society for Information Science and Technology* (Vol. 49, pp. 1–10). Baltimore, MD, USA: Wiley.

Persing, I., Davis, A., & Ng, V. (2010). Modeling organization in student essays. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 229–239). MIT Stata Center, Massachusetts, USA: Association for Computational Linguistics.

Persing, I., & Ng, V. (2013). Modeling thesis clarity in student essays. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics* (Vol. 1, pp. 260–269). Sofia, Bulgaria: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P13-1026/

Persing, I., & Ng, V. (2014). Modeling prompt adherence in student essays. In *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics* (Vol. 1, pp. 1534–1543). Baltimore, MD, USA: Association for Computational Linguistics. https://doi.org/10.3115/v1/p14-1144

Persing, I., & Ng, V. (2015). Modeling argument strength in student essays. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing* (Vol. 1, pp. 543–552). Beijing, China: Association for Computational Linguistics. https://doi.org/10.3115/v1/p15–1053

Potthast, M., Hagen, M., Stein, B., Graßegger, M., Michel, M., Tippmann, M., & Welsch, C. (2012). ChatNoir: A search engine for the ClueWeb09 corpus. In *The 35th international ACM SIGIR conference on research and development in information retrieval* (p. 1004). New York, NY: Association for Computing Machinery.

Potthast, M., Hagen, M., Völske, M., Gomoll, J., & Stein, B. (2012). Webis Text Reuse Corpus 2012. In *51st annual meeting of the Association for Computational Linguistics*. Sofia, Bulgaria: Zenodo. https://doi.org/10.5281/zenodo.1341602

Potthast, M., Hagen, M., Völske, M., & Stein, B. (2013). Crowdsourcing interaction logs to understand text reuse from the web. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics* (pp. 1212–1221). Sofia, Bulgaria: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P13-1119

Raman, K., Bennett, P. N., & Collins-Thompson, K. (2013). Toward whole-session relevance: Exploring intrinsic diversity in web search. In *The 36th international ACM SIGIR conference on research and development in information retrieval* (pp. 463–472). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2484028.2484089

Smith, C. L., & Kantor, P. B. (2008). User adaptation: Good results from poor systems. In *The 31st annual international ACM SIGIR conference* (pp. 147–154). New York, NY: Association for Computing Machinery.

Vakkari, P. (2003). Task-based information searching. *Annual Review of Information Science and Technology*, *37*(1), 413–464. https://doi.org/10.1002/aris.1440370110

Vakkari, P. (2020). The usefulness of search results: A systematization of types and predictors. In *CHIIR '20: Conference on human information interaction and retrieval* (pp. 243–252). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3343413.3377955

Vakkari, P., & Huuskonen, S. (2012). Search effort degrades search output but improves task outcome. *Journal of the American Society for Information Science and Technology*, *63*(4), 657–670. https://doi.org/10.1002/asi.21683

Vakkari, P., Volske, M., Hagen, M., Potthast, M., & Stein, B. (2018). Predicting retrieval success based on information use for writing tasks. In *Digital libraries for open knowledge: Proceedings of the 22nd international conference on theory and practice of digital libraries* (pp. 161–173). Porto, Portugal: Springer. https://doi.org/10.1007/978-3-030-00066-0_14

Vakkari, P., Völske, M., Potthast, M., Hagen, M., & Stein, B. (2019). Modeling the usefulness of search results as measured by information use. *Information Processing & Management*, *56*(3), 879–894. https://doi.org/10.1016/j.ipm.2019.02.001

Wachsmuth, H., Khatib, K. A., & Stein, B. (2016). Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: Technical papers* (pp. 1680–1691). Osaka, Japan: The COLING 2016 Organizing Committee. Retrieved from https://www.aclweb.org/anthology/C16-1158/

Wildemuth, B. M., de Bliek, R., Friedman, C. P., & File, D. D. (1995). Medical students' personal knowledge, searching proficiency, and database use in problem solving. *Journal of the American Society for Information Science*, *46*(8), 590–607.

Yu, R., Gadiraju, U., Holtz, P., Rokicki, M., Kemkes, P., & Dietze, S. (2018). Predicting user knowledge gain in informational search sessions. In *41st international ACM SIGIR conference on research and development in information retrieval* (pp. 75–84). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3209978.3210064

How to cite this article: Vakkari P, Völske M, Potthast M, Hagen M, Stein B. Predicting essay quality from search and writing behavior. *J Assoc Inf Sci Technol*. 2021;1–14. https://doi.org/10.1002/asi.24451