How Writers Search

Analyzing the Search and Writing Logs of Non-fictional Essays*

Matthias Hagen Martin Potthast Michael Völske Jakob Gomoll Benno Stein

Bauhaus-Universität Weimar
99421 Weimar, Germany
<first name>.<last name>@uni-weimar.de

ABSTRACT

Many writers of non-fictional texts engage intensively in exploratory web search scenarios during their background research on the essay topic. Though understanding such search behavior is necessary for the development of search engines that specifically support writing tasks, it has neither been systematically recorded nor analyzed. This paper contributes part of the missing research: We report on the outcomes of a large-scale corpus construction initiative to acquire detailed interaction logs of writers who were given a writing task on 150 pre-defined TREC topics. The corpus is freely available to foster research on exploratory search. Each essay is at least 5000 words long and comes with a chronological log of search queries, result clicks, web browsing trails, and fine-grained writing revisions that reflect the task completion status. To ensure reproducibility, a fully-fledged, static web search environment has been created on top of the ClueWeb09 corpus as part of our initiative.

In this paper, we present initial analyses of the recorded search interaction logs and overview insights gained from them: (1) essay writing behavior corresponds to search patterns that are rather stable for the same writer, (2) fact-checking queries often conclude a writing task, (3) recurring anchor queries are often submitted to not lose the main themes or to explore new directions, (4) query terms can be learned while searching and reading, (5) the number of submitted queries is not a good indicator for task completion.

1. INTRODUCTION

The web has fundamentally changed how writers of non-fictional texts approach their task. In the past, research on a topic and writing about it typically happened separately in time and space (e.g., research in the library, writing at home). Nowadays, both can be done more or less simultaneously, since web search engines retrieve relevant information on almost any topic. Therefore, writers cab easily switch between search and writing whenever they perceive gaps of knowledge (i.e., information needs). This situation has accelerated the rate at which non-fictional texts are written as well as significantly decreased the costs of doing so, which is particularly true in cases where the resulting texts are not expected to be award-worthy, but merely publishable.

This paper studies the writing process and search behavior of writers in action: we hired 12 authors to write a total of 150 essays on that many topics, at least 5,000 words each, while recording a fine-grained log of text revisions, search queries, result clicks, and browsing. To attain reproducibility, we chose topics from the TREC web track and set up a static web search environment based on the ClueWeb09. Our search engine employs BM25F as the retrieval model, its user interface resembles those of commercial search engines, and its performance was optimized to allow for an average retrieval time of less than five seconds. While the retrieval model of our search engine obviously does not compete with those of commercial search engines in terms of retrieval quality, that may be to our advantage since our writers had to engage with our search engine to find sufficient material to write an essay of the aforementioned length. Given budget limitations, we further attempted to shift the attention of our writers toward searching for information, rather than spending time pondering over formulations, by allowing them to reuse in their essays the texts they found in the ClueWeb09. Nevertheless, the final essays were still required to be coherent and consistent—often resulting in reformulations of copy-pasted texts. Altogether, despite the outlined limitations, we created the largest and most realistic resource available to date to study the search behavior of writers at scale. The corpus is freely available.

We present the results of an exploratory corpus analysis. After a review of related work, Section 3 describes our corpus. Section 4 then discusses analyses of various aspects of the logs and the observed writers' search behavior. We examine the structure of the submitted queries, reveal and distinguish between two elementary search strategies employed that also reflect the writing behavior, and analyze how different working phases relate to each other. Section 5 summarizes our findings and gives directions for future work.

2. RELATED WORK

Marchionini [17] distinguishes closed-ended and open-ended search tasks, where the former is a search for a particular fact, and the latter does not necessarily have a an unequivocal result. For example, compare the search for a date of birth with that for the best hotel for an upcoming trip. Open-ended tasks are often referred to as *exploratory search*

[*] Extended version of our EuroHCIR 2013 workshop contribution [23].

CHIIR '16, March 13-17, 2016, Carrboro, NC, USA
© 2016 ACM. ISBN 978-1-4503-3751-9/16/03…$15.00
DOI: http://dx.doi.org/10.1145/2854946.2854969

since they do not necessarily lead to only one correct answer but they help to build a mental model of a topic [30]. White and Roth [33] describe exploratory search as an iterative, multi-tactical process, where the user explores the information space as extensively as necessary to fulfill an open-ended information need. Closed-ended searchers may iteratively refine their queries as well, but they usually zero in on a specific, targeted piece of information. Exploratory searchers, instead, explore the information space extensively; while examining search results, they obtain clues for their next steps [35]. The challenge of exploratory search is to design retrieval models that support users in these tasks. Web search engines are typically tuned towards precision, which limits the chance of finding loosely related information. But exploratory search is more recall-oriented [17]. This can be supported via rapid query refinement in the early phase of a search [32], supporting facets such as search result clustering [28], and leveraging the searcher's context, e.g., via pseudo-relevance feedback [36]. The task of writing an essay on a given topic is open-ended and exploratory.

The query logs of search engine users are a valuable resource to study their goals. However, Kurth [15] argued early on that no measure that can be derived from user interactions alone explains the user's intentions. Researchers nevertheless rely on such measures for the lack of a better alternative. Typical measures found in the literature include the number of queries submitted by a user, the average number of terms and clicks per query, and the time between query and first click [3]. Log analyses often measure further attributes from a more global context, such as the number of physical sessions to complete a task. Machine learning algorithms then exploit a wide range of such and similar measures [1, 2, 7, 5, 20]. For example, Agichtein et al. [2] predict whether a user is likely to resume a suspended session within the next few days. After determining the dominant topic of the majority of the queries using data from the Open Directory Project, their approach is able to automatically decide for each query whether it is related to the task or not. Such approaches face the typical dilemma of machine learning, namely that the results obtained often do not reveal *why* the classifier works. User behavior is hardly ever characterized, which is the goal of our work. Also, the reproducibility of these studies is often limited by the fact that query logs cannot be shared for privacy reasons. In our case, we can safely share the entire query logs, since they are not interspersed with personal queries.

While log analyses have been conducted for a long time, exploratory search has shifted into focus only recently: to the best of our knowledge, Qu and Furnas [24] were the first to design a corresponding study. Based on the sense-making model [9, 26], they studied the relation between information seeking and construction of a mental representation. In this regard, not only the interactions of their 30 participants with the search system were recorded, but participants were also asked to prepare an outline for a 1-hour talk. Interestingly, Qu and Furnas found that the resulting talk structure strongly correlated with that of the participant's bookmark folders. Human judges rated the topical similarity between consecutive queries and assigned each query to one of the bookmark folders. Qu and Furnas visualized this information on a timeline to show when which query occurred, which folder it referred to, and which web page was bookmarked in this context. The visualizations for all 30 subjects reveal the influence of emerging structure on the following search. Moreover, 14 out of 30 participants used their folder structure as a roadmap for subsequent search. The authors conclude that search engines should support users, for instance, by analyzing the structure of their bookmark folders.

Egusa et al. [11] pursued a similar approach asking 35 undergraduate students to produce a concept map of their understanding of a given topic before and after searching. A concept map is a graph consisting of named entities and labeled connections between them. By analyzing the differences between the before- and after-maps, Egusa et al. develop a new task performance measure for exploratory search tasks. Such a measure goes beyond traditional IR measures in assessing not only precision but also the *benefit* a user has from a set of search results [33]. In this regard, Vakkari [30] differentiates between evaluating *search engine output*—the precision of a result list with regard to the submitted query—and *task outcome*, which describes how well the system supported the user in fulfilling the task. A high precision does not necessarily lead to good overall task performance. Egusa et al. performed their experiments on only two very broad (i.e., open-ended) topics, namely "Politics" and "Media." The task was to find and compare different opinions about these topics. The before- and after-maps were analyzed with respect to the number of kept, discarded and inserted nodes, links and labels. Among other findings, nearly as many deletions as insertions occurred. This indicates that people not only gather new information while exploring a topic but also adjust their existing knowledge. However, the authors conclude that applying descriptive statistics on concept maps cannot serve as a measure for the performance of an exploratory search system. They argue that one has to conduct more qualitative analyses of the described concepts and users' searching behavior.

Vakkari and Huuskonen [31] designed a study that concentrates on the search process, especially the effort that users put into the search, and how it is interlinked with the task outcome. Within the scope of a term's course, medical students were asked to find information with a domain-specific search engine in order to write an essay on a medical topic. The search log interactions were examined with respect to the applied search tactics (narrowing and broadening of queries, use of logical operators, etc.) and effort variables (like number of sessions or the number of read, but not cited articles). The essays' grades as awarded by the course's instructors were used as a performance measure for task outcome. Vakkari and Huuskonen show several interesting relationships between search process, output and outcome variables. They report a negative correlation between diversity of queries, search engine precision, and essay scores: the broader the queries were formulated, the lower the system's precision, yet the higher the essay scores. A very similar correlation was observed for search effort: the more sessions a student needed to write the essay, the lower was the system's overall precision because of the larger result set, but the higher was the quality of the essay.

Liu and Belkin [16] investigated the association between newspaper article writing and information search in a study with 24 undergraduate students. The participants worked on one of two writing tasks, with intermediate stages of task completion recorded at the end of each of three sessions. As such, this study approaches what is possible with our own log, albeit at a much lower level of granularity.

Our efforts to construct a corpus for exploratory search have been guided by the aforementioned approaches, addressing several shortcomings: *(a) Task diversity.* Qu and Furnas [24], Egusa et al. [11] as well as Liu and Belkin [16] employed only two different topics. Vakkari and Huuskonen [31] employ eleven topics, but all of them from the medical domain. We employ 150 topics, derived from the TREC web track, which are diverse and can be understood by laymen. *(b) Connection of search and task outcome.* Qu and Furnas [24] and Egusa et al. [11] do not provide revisions of task outcomes. Our study aligns all search interactions with text revisions on a time line, which allows fine-grained analysis of the connection between search and task outcome, as proposed by Järvelin et al. [14]. *(c) Experimental setup and reproducibility.* Qu and Furnas [24], and Liu and Belkin [16], asked participants to use a search system in their lab—a maximally obtrusive setting [13]—whereas our participants could work from home. Unlike the other studies, we employ a well-known web corpus frequently used for evaluation purposes to create a static search scenario, that can be reproduced even after years. *(d) Incentives and motivation.* All four studies recruit undergraduate students as study subjects, which often introduces bias with regard to diversity and motivation. Vakkari and Huuskonen ensure proper motivation, since their participants were graded and had the chance of earning credit points by completing the course; Liu and Belkin's participants received monetary compensation. In our case, we hired (semi-)professional writers from all over the world with a diversity of backgrounds, we had them sign a contract, and paid them on an hourly basis.

3. DATASET DESCRIPTION

Our Webis Text Reuse Corpus 2012 (Webis-TRC-12)[1] consists of fine-grained interaction logs for the writing of 5,000 word long essays on 150 different topics—147 essays with a few pre-defined documents that the authors should use, 150 essays for which the authors were actually asked to search for relevant information. The dataset consisting of the latter 150 essays and their writing process (i.e., the so-called "search" subset of the Webis-TRC-12) forms the basis of the analyses in this paper. Each topic was derived from the TREC Web track topics of the years 2009–2011; reformulated to result in an essay writing task on the topic instead of information finding only. Since most topics are rather broad, the authors had to resort to exploratory search for their research using the ClueWeb09 search engine ChatNoir [21] based on the BM25F retrieval model [25]. All search interactions are logged in the Webis-TRC-12 alongside the revisions of the actual texts. A new revision was added whenever the authors stopped typing for 300ms in our online editor provided for their essay writing. This way, the dataset allows to complement analyses of the task progress in form of essay completion with a fine-grained search behavior log. The authors were allowed to reuse passages from the found web documents but instructed to indicate the sources. The whole process of the essay revision logging and the framework involved was discussed in much detail in our previous publication focusing on the writing behavior [22]. In the paper at hand, we focus on the search interactions to gain insights on how writers search. To be rather self-contained, we give some key figures on the essays first but then concentrate on the search log. The dataset is freely available for use by other researchers.[2]

3.1 The Essays in the Webis-TRC-12

The outstanding property of the essays compared to other corpora is that very fine-grained intermediate states from the beginning up to essay completion are synchronized with the writer's search behavior. Most essays are around 5,000 words long as requested—there are only two shorter ones due to difficulties in finding useful documents in the ClueWeb09. For example, one author should write about the HP Mini 2140 notebook but since its market launch falls in the crawling period of the ClueWeb09 [6] only few announcements of the product could be found. About half of the essays contain reused text from 11 up to 21 different ClueWeb09 documents, only one fourth of the essays contain less than 11 sources, with a minimum of only 3. Section 4.5 will later show that writers with many sources can be considered slightly more dedicated to the task than writers with only few sources. For more details on the corpus creation process, we refer to our previous publications [22, 23].

3.2 The Search Log of the Webis-TRC-12

The search log consists of 150 files, each containing all search interactions for one essay. There are three different types of interactions in the log: (1) queries submitted by the user along the shown ranked results including snippets; (2) document views, characterized by the visited URL and a type (click on a search result, a trail click, or a revisit via some bookmark);[3] (3) revision numbers of text-writing interactions, serving as a cross reference to the actual essays.

Each interaction has a timestamp and an anonymized IP address, which may give a clue about different workstations that writers worked with. Documents are referenced by both their ClueWeb09 IDs, as well as their real URLs.

Table 1 summarizes important statistics on our dataset. The search log contains 13,609 queries, 16,698 document views and 6,123 text-writing interactions by 12 different authors. All interactions took place in about half a year, spanning 166 days. The longest time period spanned by one essay is 56 days, yet the author actually worked on only 12 of these days and paused work on the other 44 days. The majority of authors worked for about 6 days on a topic before essay completion, whereof 5 days involve actual working phases and 1 day involves no working hours at all (median values).

When separating all interactions into physical sessions with a cut-off time of 15 minutes—i. e., an author is considered to be inactive after a gap of 15 minutes—the log contains 2,797 physical sessions, resulting in an average of 18.6 sessions per essay or 3.4 sessions per working day, respectively. The shortest sessions only span a couple of minutes and often involve only a few edits, whereas the longest sessions last up to 253 minutes (more than four hours). A more detailed analysis of these sessions and a visualization scheme is provided in Section 4.2.

A rather surprising number is the ratio of unique queries among all submitted queries, which is about one quar-

[1] The name's inspiration is the utilization of the final essays in the PAN lab's shared task on text reuse detection.

[2] http://webis.de/corpora

[3] Result clicks are views of search results. Followed links in such documents form the trail clicks. All other document views are bookmark clicks when the document has been visited before; otherwise, they are categorized as "unknown."

Table 1: Key figures of searching and writing for all essays in the Webis-TRC-12 "search" subset.

	Min	Q1	Mdn	Avg	Q3	Max	Sum
Queries							
– per essay	4.0	40.0	68.0	90.7	117.0	612.0	13,609
– per essay (unique)	1.0	12.0	20.0	23.6	31.5	121.0	3,538
– per physical session	0.0	0.0	0.0	4.9	4.0	231.0	13,609*
Clicks							
– per essay	12.0	55.0	87.0	111.3	144.5	431.0	16,698
– per essay (unique)	8.0	44.5	67.0	74.5	101.0	259.0	1,1181
– per physical session	0.0	0.0	1.0	6.0	6.0	164.0	16,698*
– per query	0.0	0.0	0.0	0.2	0.0	76.0	8,779
Clicks per essay							
– on results	5.0	30.5	49.0	58.5	75.5	280.0	8,779*
– trail clicks	0.0	13.5	33.0	52.8	73.0	332.0	7,919
Writing sessions							
– per essay	11.0	28.0	42.0	46.3	59.5	178.0	6,943
– revisions (thousands)	0.2	1.8	2.9	2.9	3.8	6.8	–**
– words (thousands)	0.7	4.8	5.0	5.0	5.2	13.9	–**
– paste events	0.0	13.0	25.0	28.6	39.0	134.0	4,291
– references	3.0	11.0	16.0	18.4	21.0	69.0	2,761
Work time per essay							
– days passed	1.0	4.0	6.0	8.6	9.0	56.0	–**
– working days	1.0	4.0	5.0	5.5	7.0	17.0	–**
– working hours	1.8	5.2	7.5	7.9	9.8	23.0	1,191
– physical sessions	2.0	11.5	16.0	18.6	23.0	55.0	2,797
Minutes spent							
– reading per click	0.0	0.1	0.4	0.7	0.8	15.0	11,236
– writing per session	0.0	0.5	2.2	7.4	8.9	145.2	51,126

*Equal to some above value by definition.
**Sum not given to avoid misinterpretation.

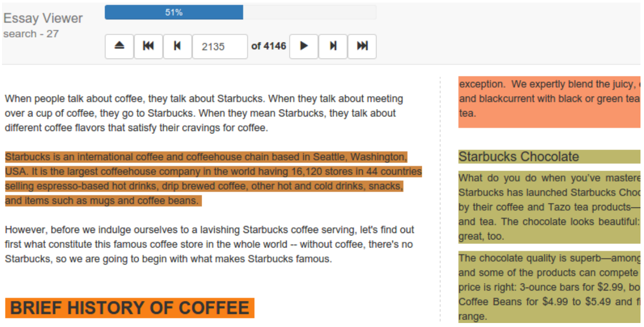
ter (3,538 to 13,609). About half of this effect is explained by the interface of the search engine: Writers were shown the top ten results first, and could request 100 results by clicking a "more"-button. This accounts for 6,874 queries with 10 results and 6,727 follow-up queries requesting 100 results. Almost always, the authors clicked "more." A further explanation why there is only about one fourth of unique queries is provided in Section 4.3: many authors submit identical queries in different sessions or even in a row.

It is also remarkable that more than half of the physical sessions contain no query submission. In fact, the third quartile of queries per session is 4, meaning that 2,097 sessions contain $\leq$ 4 queries. Almost all queries (12,094 of 13,609) were submitted in 700 sessions only. The document views are similarly distributed: 14,421 of 16,698 views take place in only 700 sessions. This indicates that text-writing interactions form the largest part of most physical sessions.

A closer inspection of the statistics reveals two other interesting aspects. First, the number of result and trail clicks is quite balanced, which indicates that the writers genuinely followed exploratory search strategies and not just entered look-up queries. Second, the writers did not spent too much time reading the clicked documents. With a median value of 0.4 minutes ($\widehat{=}$ 24 seconds) and a third quartile of 0.8 minutes ($\widehat{=}$ 48 seconds) only few clicked documents seem to be worth reading in-depth. For example, only 661 documents, which is about 4% of all clicks, were viewed for two and a half minutes or longer. One reason could be that the writers just copy-pasted content from some of the results and only read the content while editing it in the essay editor.

4. WRITERS' BEHAVIOR

The following analyses of our dataset aim to shed some light on how to characterize and understand the writers' behavior in exploratory search tasks.



When people talk about coffee, they talk about Starbucks. When they talk about meeting over a cup of coffee, they go to Starbucks. When they mean Starbucks, they talk about different coffee flavors that satisfy their cravings for coffee.

Starbucks is an international coffee and coffeehouse chain based in Seattle, Washington, USA. It is the largest coffeehouse company in the world having 16,120 stores in 44 countries selling espresso-based hot drinks, drip brewed coffee, other hot and cold drinks, snacks, and items such as mugs and coffee beans.

However, before we indulge ourselves to a lavishing Starbucks coffee serving, let's find out first what constitute this famous coffee store in the whole world -- without coffee, there's no Starbucks, so we are going to begin with what makes Starbucks famous.

BRIEF HISTORY OF COFFEE

exception. We expertly blend the juicy,
and blackcurrent with black or green tea
tea.

Starbucks Chocolate

What do you do when you've mastere
Starbucks has launched Starbucks Choc
by their coffee and Tazo tea products—
and tea. The chocolate looks beautiful:
great, too.

The chocolate quality is superb—among
and some of the products can compete
price is right: 3-ounce bars for $2.99, bo
Coffee Beans for $4.99 to $5.49 and f
range.

Figure 1: Screenshot of the essay viewer interface.

4.1 Visualizing Essay Writing

In order to better understand the essay writing process, we implemented a web application that shows all revisions of a given essay in sequence. A screenshot of the interface is shown in Figure 1: The controls at the top of the page allow stepping forward and backward through the essay revisions, or jumping to a specific revision. The rest of the page shows the current state of the essay. Different colors indicate different ClueWeb09 sources for copied or paraphrased text. We envision extending this tool to include information from the query log, such that queries occurring at a given point in time can be correlated with their contemporary writing interactions. The tool is available alongside the corpus.

4.2 Visualizing Writer Interactions

In a visual illustration of the logged interaction, we examine the temporal course of actions that the authors took during their essay writing task. To this end, the physical working sessions are determined based on a 15 minutes inactivity gap. However, only the text-writing interactions have an exactly known end time; for query and click interactions, we estimate the durations. For queries, we apply a threshold of 60 seconds, because we assume that a writer would not stare on the result list for more than one minute without clicking any result. For clicked documents, we estimate the reading time based on the document length and an assumed reading speed of 250 words per minute [8]. A solid (green) line further shows the development of essay length in the sessions. Figure 2 shows examples of three topics—visualizations for all topics are available alongside the corpus. Each row depicts a physical session, and the horizontal dashed lines divide different working days (most of the sessions in the plots are about one hour). The beige blocks represent text-writing interactions, and the blue and red ones depict queries and document views, respectively.

The author of the essay on topic 29 submitted rather few queries but seems to have worked very purposeful. Writing often directly follows document views and it seems that the author deliberately decided to learn and write about some particular aspect and visited a couple of documents in order to collect the needed information. In contrast to this, the author of the essay on topic 27 has a very different working style. Starting with a couple of sessions in which the author foraged all possibly needed information, almost all sessions from the third working day on deal with rewriting and removing content from the priorly collected sources. Following our previous notation [22], we call the first type a "build-up" writer and the second a "boil-down" one. In Section 4.4, we will also relate this to different searching behavior.

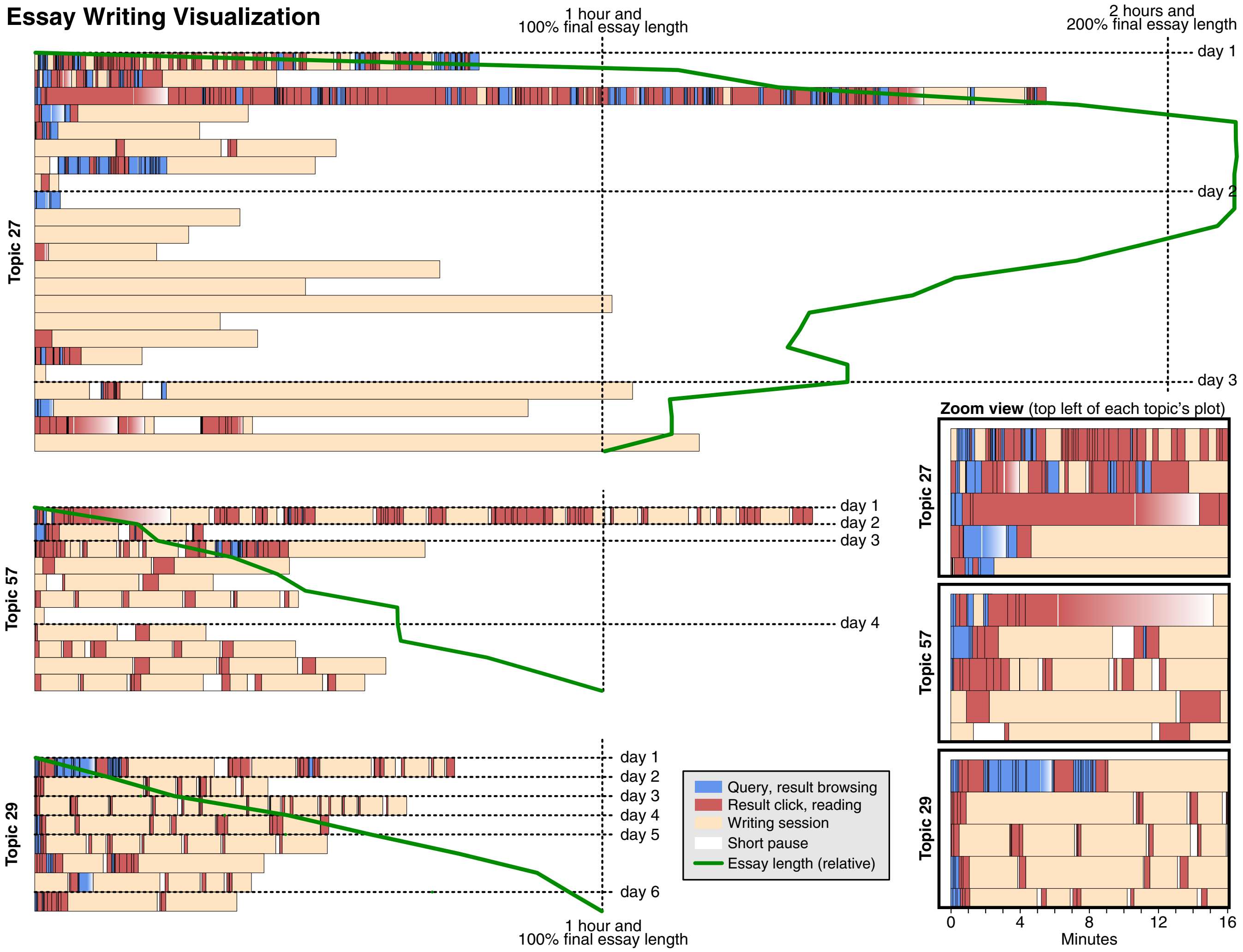


Figure 2: Visualization of the interactions for a selection of three of the 150 essays. Each stacked bar denotes an uninterrupted working session (15 minutes inactivity gap). Bar lengths indicate work time in minutes, blue boxes indicate querying and result browsing, red boxes indicate document views, and beige boxes indicate writing. White boxes and gradients in red or blue boxes indicate short pauses. The solid green line denotes the current essay length relative to the final essay.

There is another interesting detail about the essay on topic 27: In the session before the last, a couple of document views are followed by very short writing interactions that influence the essay length only marginally. This can be observed for many topics and different authors and was also recognized by Vakkari et al. [30]. One explanation could be that writers check their essay for possibly missing but important text passages from priorly selected sources or that they double-check the facts in their essays.

4.3 Query Formulation

In exploratory tasks searchers learn and extend or adapt their knowledge about a topic [11]. We expect the queries for an essay to also develop over time and examine when in the process specific terms occur and where they might stem from. For each query term entered for the first time, we assign it to one of the possible origins: the task description, a previously clicked document, the title or snippet of a previously shown search result, or the writer's initial knowledge. If a term has not occurred during any of the prior interactions, it is classified as *prior knowledge* only.

Table 2: Origins of learned query terms.

Prior	Task	Search Results		
knowledge	**description**	Title	Snippet	Clicked doc.
312 (8.4 %)	902 (24.3 %)	291 (7.8 %)	1,067 (28.7 %)	1,147 (30.8 %)

Table 2 shows the origins of all 3,719 distinct query terms that appeared in the queries for the 150 topics. Almost all terms could potentially be learned during work on the topic. Figure 3 (middle and right) shows when in the search process a writer introduced new terms and where they are likely to come from for topic 29 and 133. On the x-axis, one can see the current query number. The y-axis displays all clicked documents, and the staircase-shaped line depicts which click(s) happened as a result of which query. For topic 29, the first three clicks happened for the sixth query, another click followed after submitting the ninth query, and so on. The dots indicate a new term in the query and all previously clicked documents that contain this particular term.

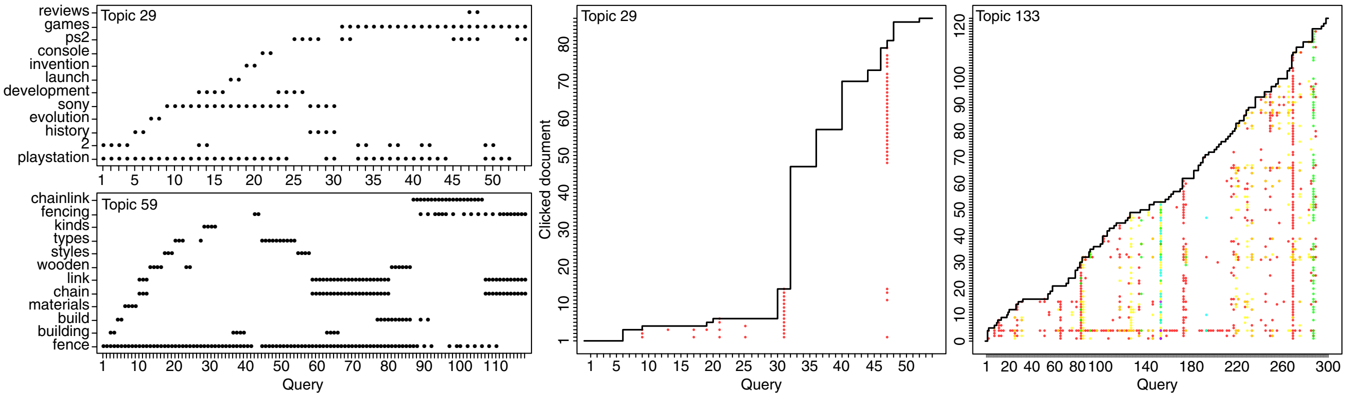


Figure 3: Query composition for topics 29 and 59 (left). Learned terms for topic 29 (middle) and 133 (right).

If two or even more terms were introduced in only one query, each of the terms is represented by another color. For instance, in topic 29 the queries 31 and 47 introduce the terms "games" and "reviews." These terms were contained in almost all of the clicked documents that were visited before the respective queries. Such a vertical line of dots can be interpreted as a change of subtopic because the writer ignored an often occurring term for quite a long time and then decided at some point to finally search for it. About 70 of the topics contain such clear subtopic changes based on recently visited documents. Topic 133 also shows another interesting pattern: a horizontal line (Figure 3 (right)). This indicates that document number 4 was influential for many queries (it is a detailed overview on the Declaration of Independence, the main theme of the topic).

With respect to what terms were used in which queries, also the left part of Figure 3 visualizes usage: the terms on the y-axis and the queries on the x-axis. From the two example topics it is obvious that many queries have numerous identical, immediate follow-up queries. Half of it can be explained through clicks on the "more"-button requesting 100 instead of 10 results. Often a query is submitted more than twice when there was a session break in between and the writer started with the same query again. However, there are also some oddities like the query `chain link fence` that is submitted ten times in a row for topic 59 (queries 67 to 76 in topic 59). We have no satisfying explanation for this behavior; maybe the search engine was slow at this time such that the writer submitted the query again before having seen any result.

We consider the identical queries that are submitted from time to time to be *anchor* queries. The results of such a query can point to many directions for further investigations and a writer might return to this query as soon as the work on one subtopic is finished. Second, anchor queries can serve to keep track of the main theme at any time and keep the writer on course. And third, writers might bring recently acquired knowledge into line with older knowledge structures and therefore want to return to previously seen documents. Typical anchor queries for many topics reflect the main theme of the task (i.e., the TREC topic itself).

4.4 Search Strategies

We now focus on elementary differences in writers' searching strategies. Figure 4 shows the extreme cases of submitted queries over essay revisions for four authors (axes normalized to percentages). The curves are organized to highlight the spectrum of different search behavior for individual authors. Authors 2, 5, and 21, for instance, have topics for which they submit most of the queries rather early, but also topics with most queries at the end only (i.e., probably fact checking). Typically, sets of queries are submitted in short "bursts," followed by extended periods of writing, which can be inferred from the plateaus in the curves. For author 7, all the topics show a more linear increase of queries over the whole writing time for all topics, indicating continuous switching between searching and writing. From these observations, it can be inferred that query frequency alone is not a good indicator of task completion or the current stage of a task even within a single author. Moreover, exploratory search systems have to deal with a broad behavior spectrum and be able to make the most of few queries, or be prepared that writers interact only a few times with them.

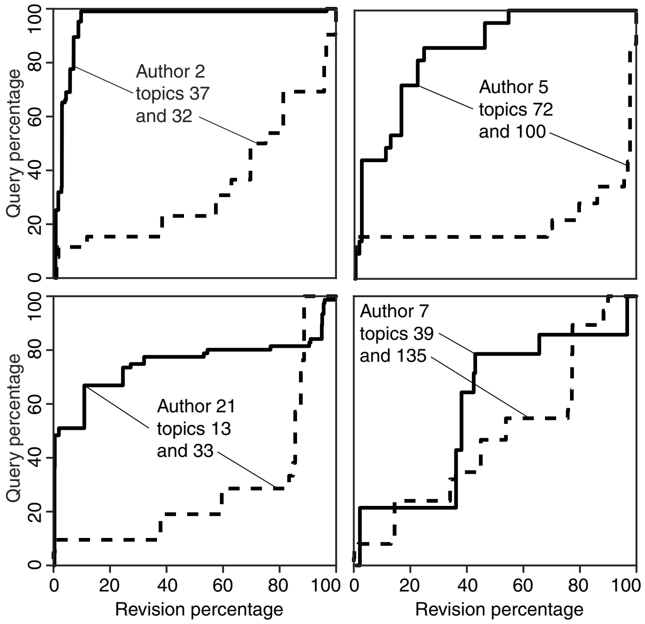


Figure 4: Examples for the spectrum of writer search behavior. Each curve shows the percentage of submitted queries (y-axis) per percentage of essay revisions (x-axis). For each author, we show the topics with the largest and smallest area under the curve (i.e., early queries vs. late queries).

To further distinguish search behavior, we focus on the number of queries and clicks. As observed in Section 3.2,

Table 3: Median values for essays of clickers and queriers (Mann-Whitney U-test).

	Clickers	Queriers	Significance of difference
Queries	47.0	107.0	$U = 1058.5,\ z = -6.148,\ p < 0.01$
Clicks	102.5	79.5	$U = 2074.0,\ z = -2.136,\ p < 0.05$
Pastes	39.0	19.0	$U = 1361.5,\ z = -4.952,\ p < 0.01$
References	17.5	15.0	$U = 1876.0,\ z = -2.921,\ p < 0.01$

some authors submit only few queries but follow long click trails; others submit a variety of queries but rarely click on search results. We call the authors following one of these two strategies *clickers* and *queriers*. To distinguish between clickers and queriers, we count the number of queries and clicks that are performed until a document is clicked that is also used in the essay. It is not important how many queries and clicks occurred overall but only how many of them occur between two clicks on such reference documents. The analysis for all essays reveals the two groups among our authors. Authors 5, 7, 20, 21 and 04 are clickers, and the authors 2, 6, 17 and 18 are queriers. Authors 1, 14 and 25 have worked on at most two topics only, yet the trend shows that they tend to be clickers.

Table 3 highlights the differences between clickers and queriers. Except for the number of clicks, which is also fairly high for queriers, all differences between both groups are highly significant as shown by a Mann-Whitney U-test (the data is not normally distributed). The fairly high number of clicks in the querier group simply seems to depend on the number of queries submitted. After all, the distributions of clicks for both groups differ not as much as the distributions of queries, pastes and references. This underpins the assumption that writers in exploratory search tasks consume some informative content before considering themselves to have learned enough. It is notable that clickers paste about twice as often as queriers do. It seems plausible that clickers pick up several possibly useful text passages during their information exploration phase, which they retain in their essays for later use. The number of used references confirms this trend and it can be stated that queriers seem to be more selective with their reference documents than clickers.

4.5 Writer Dedication

Besides different search strategies, we also want to explore whether our data allows us to measure the degree of *writer dedication* to the exploratory search task. We try to reflect writer dedication by the effort a writer puts into the treatment of the task, which can be a valuable information for a search engine. For example, a truly dedicated writer might be interested in additional resources beyond the original query, whereas a writer who works only unwillingly on a task like essay writing might be only interested in overview pages without too many details. Recent studies investigating user engagement [18, 19] go beyond the simple features we can explore below, but we think that our search log-derivable measures can still be useful.

To distinguish "lazy" from more dedicated writers, we use the following nine features per essay: number of distinct queries, number of distinct clicks, number of copy-paste interactions, number of used references, total working hours, time spent for reading documents, time spent for writing, number of physical sessions, number of handled subtopics (determined by the number of session IDs a search mission detection algorithm returned [12]). In a next step, a ranking of all topics is produced for each feature individually, and each essay gets a score depending on its rank. For example, the essay on topic 133 contains the most distinct queries and thus obtains 121 points (it is not 150 because 29 essays share the same number of distinct queries and obtain the same score). For the feature "distinct clicks," the essay on topic 133 is only on rank 18 and obtains 77 points. This is done for all features and the scores are summed up per essay; the resulting ranking is shown in Table 4. Remarkably, nine of the top-10 essays were written by author 2, who seems to have worked with high dedication on many essays, whereas authors 6 and 24 seem to have worked with little enthusiasm—even though the authors picked their favorite topic from the remaining ones when starting a new essay.

Table 4: Essays with topic (T) and author IDs (A) ranked (R) by the writer dedication score (S).

R	T	A	S	R	T	A	S	R	T	A	S
1	58	2	551	51	150	24	334	101	73	24	201
2	53	2	538	52	138	2	331	102	81	17	200
3	110	2	524	53	57	5	330	103	24	14	196
4	13	21	523	54	36	5	326	104	100	5	196
5	67	2	503	55	48	18	323	105	66	20	194
6	27	2	499	56	50	2	320	106	102	24	194
7	49	2	498	57	117	2	320	107	69	24	191
8	144	2	493	58	55	21	319	108	126	6	191
9	10	2	484	59	137	5	317	109	14	5	189
10	22	2	479	60	65	17	314	110	40	17	188
11	133	17	476	61	47	2	313	111	15	20	186
12	80	2	470	62	1	17	311	112	94	17	184
13	88	2	469	63	63	5	311	113	90	18	178
14	51	2	468	64	107	17	308	114	95	5	178
15	139	5	467	65	25	17	304	115	83	18	173
16	45	21	466	66	92	18	304	116	4	18	170
17	37	2	455	67	115	5	301	117	103	20	169
18	71	21	448	68	12	5	298	118	20	5	168
19	127	2	448	69	39	7	296	119	140	18	165
20	86	21	446	70	105	7	295	120	85	17	163
21	42	17	444	71	64	2	291	121	34	18	162
22	8	2	441	72	75	2	289	122	46	7	159
23	120	21	430	73	99	7	285	123	16	18	155
24	141	2	422	74	109	7	282	124	148	20	155
25	106	21	417	75	125	21	279	125	72	5	152
26	17	2	414	76	60	18	276	126	101	24	152
27	82	2	414	77	145	17	273	127	104	7	150
28	98	21	406	78	19	20	267	128	9	17	149
29	87	17	404	79	54	6	263	129	142	20	147
30	11	24	403	80	30	2	262	130	136	7	139
31	114	5	399	81	41	7	252	131	61	18	135
32	59	2	394	82	77	5	252	132	129	6	131
33	76	21	394	83	35	5	248	133	123	1	127
34	5	17	393	84	118	25	248	134	84	18	126
35	70	20	392	85	6	17	247	135	132	24	126
36	74	2	389	86	29	5	246	136	91	20	125
37	96	18	383	87	121	17	246	137	113	5	125
38	119	2	378	88	131	7	243	138	112	18	122
39	135	21	376	89	78	5	235	139	130	24	117
40	31	2	375	90	149	17	235	140	38	18	116
41	26	1	374	91	62	17	233	141	89	7	115
42	128	2	372	92	122	2	233	142	32	2	113
43	18	2	366	93	97	6	226	143	3	24	111
44	2	17	357	94	56	18	220	144	124	18	104
45	7	7	355	95	79	24	218	145	23	24	89
46	44	18	355	96	28	18	216	146	147	6	74
47	33	21	354	97	143	17	213	147	116	6	63
48	93	17	344	98	52	18	207	148	43	20	62
49	108	17	342	99	134	17	205	149	146	6	45
50	68	24	336	100	111	18	202	150	21	24	40

To identify the most and the least dedicated writers, we simply compute the average for each writer in order to bypass the different numbers of treated topics. It turns out that author 2 indeed belongs to the most dedicated writers with an average score of 403.5 but is slightly outperformed by author 21 with an average of 404.8. Note that author 2 worked on 33 different topics and the range of scores is distributed, whereas author 21 worked on only 12 topics, which all achieved quite high dedication scores. The least dedicated writers in our collection are author 6 and author 20 with an average score of 141.9 and 188.6, respectively. Note that the dedication ranking does not imply any conclusions on the quality of the essay itself but only about the effort that the authors spent for writing. The quality of the essay has to be determined in a separate step—an idea could be to run the essays through text reuse detection software and assign higher quality scores to essays from which the ClueWeb09 sources cannot be really detected anymore, similar to the source-based writing analyses of Sormunen et al. [27]. This could then also be used to confirm previous findings on how effort correlates with the task outcome [31].

4.6 Searching and Writing Styles

In our previous study focusing on the writing process, we found two different writing styles: build-up and boil-down [22]. The first is characterized by a rather continuous lengthening of the essay over the whole period of writing while the second style is characterized by a first quick length growth and subsequent reorganization and shortening. The essay on topic 27 reflects a typical boil-down writing while the essays on the topics 29 and 57 are build-up essays (cf. Figure 2). In our writing style study, we characterized 65 build-up essays, 65 boil-down essays, and 20 that mix both styles by a manual visual inspection [22]. Here, we now compare the writing style (essay length growth) to the search and copy-pasting behavior. The hypothesis is that in build-up essays text passages are copy-pasted in rather regular intervals (and almost immediately adapted to fit into the essay structure) while in boil-down essays most of the background research is hypothesized to happen at the beginning and thus most copy-paste interactions are to be expected at the beginning of working on a task.

As a simple measure of the search behavior, we use the regularity of copy-paste events over the course of the writing process. One could argue that queries are a better search behavior measure but with the copy-paste events we focus on the search and web interactions that actually led to some change in the essay. As for the regularity, we count the number of revisions between each consecutive pair of copy-paste events and compute the observed variance. For example, a 50-revisions essay with paste events in the revisions 10, 22 and 40, would result in the list $\langle 10, 12, 18, 10\rangle$ (also containing the revisions prior to the first and after the last paste). A low variance in this list means that the paste events are rather equally distributed over the essay revisions, whereas a high variance indicates that a writer pasted very irregularly.

As a measure for the development of essay length, we check whether at least one full word was added or removed for all subsequent revision pairs. If either is the case, a respective counter is increased. Note that for simplicity we do not count how many words have been added or removed; only the trend matters (i.e., how many revisions lengthen the essay vs. how many shorten it). In an example 50-revision essay, this might result in 20 revisions in which content was removed and 30 in which content was added. The essay thus tends to grow, as 60% of the revisions lead to a longer essay. Yet, naturally each of the essays has to grow in total to reach a 5,000-words length. Therefore, a low value like 60% rather is an indication of a boil-down writing style.

Figure 5 shows the plot resulting from the essay length development and the paste regularity for each topic. Different symbols (and colors) indicate different authors, thus revealing trends for each author's writing style. The x-axis ranges from about 50% to almost 100 %; the essays more to the right are from the more lazy writers that hardly ever rephrased something they copy-pasted. The two authors 6 and 20 who are isolated from all other authors by reaching an essay growth of ≥85% also are the least dedicated authors in Section 4.5. Many authors range in a 70% to 85% essay growth showing a build-up pattern in our earlier observations, while most essays with a growth below 70 %, here especially those by author 2, are those that show the boil-down pattern. Yet, as can be seen on the y-axis, even a boil-down pattern might come with rather regular paste events (low variance with high regularity is on the top of the y-axis) meaning that some authors boiled down individual fragments rather than all useful passages at once. Interestingly, different authors' essays form clusters in our plot contrasting search behavior with the writing progress (copy-paste regularity vs. essay growth). Knowing to which category a writer belongs can help the search engine to better tailor its results. For instance, later follow-up queries are likely for build-up writers. The search engine could take some time while the author is writing to already prepare appropriate results in a slow search fashion [29].

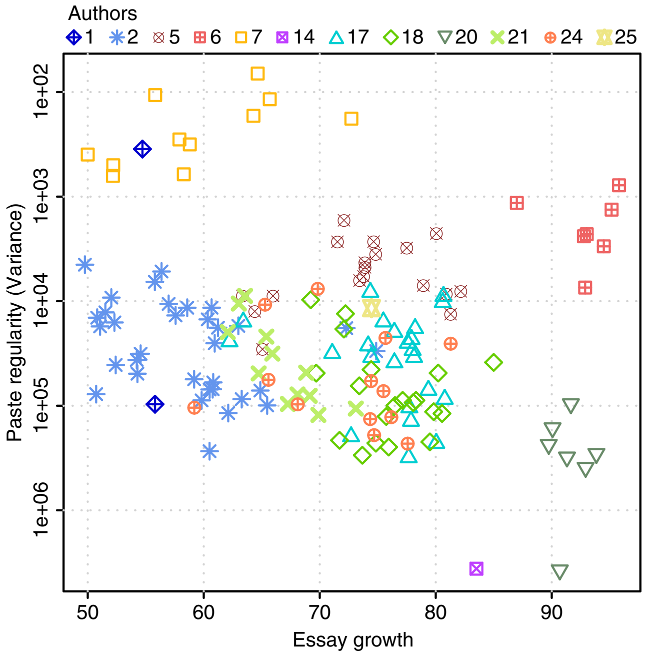


Figure 5: Authors' searching and writing style in form of the essay growth (x-axis, percentage of revisions of an essay that lengthen it) vs. the regularity of copy-pasting content from search results (log-scaled y-axis, variance of the copy-paste revision number differences, low variance = high regularity). Each essay is a data point; essays from the same author typically have similar characteristics.

4.7 Comparison of Working Phases

Finally, we investigate whether the authors work in distinct phases. Do they submit more queries early? Does writing form the major load at the end? Any patterns may inspire ideas to support writers in their respective working phases. In the beginning, a search engine could present not only results for the submitted query but also suggest shortcut queries [4] that helped other users finding relevant documents on the treated topic. While this is helpful to quickly acquire an overview of different aspects of a topic, it might not be desirable in a later phase in which a writer is only interested in specific details or just checking some facts.

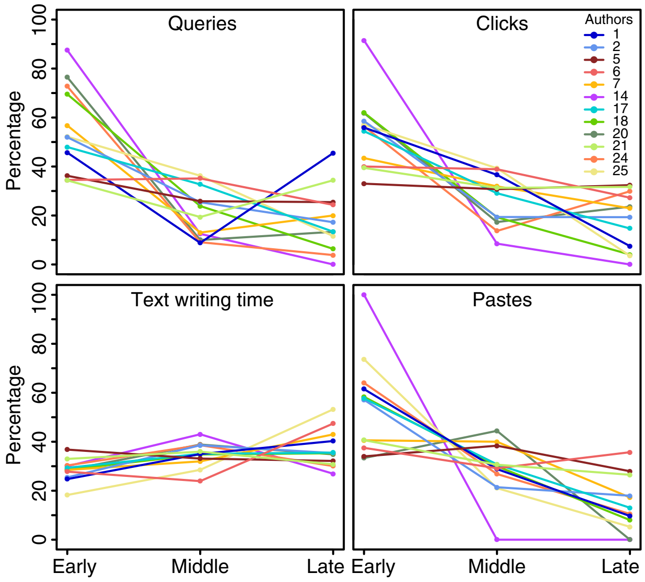


Figure 6: Work load in different working phases for all authors.

For the sake of simplicity, we subdivide each topic into three working phases—early, middle and late—by splitting up the interactions in the actual working time into three parts of equal duration. For each phase, we measure the percentage of queries, clicks, writing and copy-paste interactions that happened in that phase. For example, if 25 queries out of 50 appeared in the very beginning, the query dimension score is 50% for the early phase etc. For each author and each phase, we take the median value over all their essays to "average" the scores. Figure 6 shows the plots for all authors. The general trend is that most queries, clicks and paste interactions happen in the early phases while writing in general seems to happen more in the later phases. This is not too surprising given the fact that most authors wrote essays on topics they were not familiar with and had to search for useful content to first explore the structure of the information space [33]. Still, some authors (and even more essays) show a V-pattern in their query or click load indicating that a large portion of queries also was submitted in the last phase (e.g., authors 1, 7, and 21). Interestingly, these authors did not have a high paste load in the last phase. This indicates that the authors might have checked the essay for possibly missing text passages from previously clicked documents or that they fact-checked some of their content before completion. Interestingly, for most authors the percentages of clicks and pastes over the different phases approximately correlate. At first glance, this might indicate that the authors did no improve their precision (i.e., clicks vs. found relevant content in form of copy-pasting) over the time spent on the topic. However, an in-depth analysis of this issue is left for future work.

5. CONCLUSION AND OUTLOOK

To examine user behavior in exploratory search tasks, we analyze the search interaction logs of authors writing a 5,000-word essay for which we have a fine-grained revision history. We consider our results to constitute one more step towards understanding exploratory search and building an ideal search engine that fulfills the user's needs in such situations. Since our corpus is freely available and is related to widely used resources like TREC topics and the ClueWeb09 web crawl, replicability is ensured and comparisons can foster research on exploratory search.

In order to analyze the behavior, we propose a visualization scheme that provides a fast and easily graspable overview of all interactions throughout the writing of an essay. Although being informative for any single essay, it is difficult to draw general conclusions about user behavior just from visualizations. To this end, we conduct analyses of our dataset with respect to the search behavior complementing our previous observations on the writing process [22].

As for the querying, we find many writers submitting identical rather general "anchor" queries from time to time while working on an essay. Reasons might be to guide the exploration of the information space, to keep track of the main essay theme, or to bring recently acquired knowledge into line with earlier knowledge structures.

As for the overall search strategy, we identify two types: the *clickers* and the *queriers*. Clickers tend to visit more results and follow long click trails, whereas queriers submit significantly more queries, and often click on only few results. However, both can be very dedicated to the task.

In our analysis of writer dedication, we rank the different essays based on several features. The number of clicks, the overall reading time and the number of copy-pastes are the most discriminating features for writer dedication in our setting. Since we do not have quality assessments for the written essays yet, we did not correlate writer dedication with the quality of the essays. However, contrasting dedication with the analysis of how well automatic text reuse detection systems identify the ClueWeb09 sources of the reused passages could be a promising way in that direction. The underlying hypothesis is that easy-to-detect sources probably were not rewritten that much indicating a lower essay quality. Such writers can be considered rather "lazy" or less dedicated than writers of essays with fewer detected sources.

As for the general search behavior, we find a relationship to the build-up and boil-down writing styles [22]. By a thorough grounding on a machine derivable score for essay growth, we could relate the author types to the regularity of copy-pasting from the search results. The resulting plot shows that authors rather stick to their habits of paste-regularity and writing style over different essays. However, interestingly most authors had differences in their search behavior close to essay completion. For some essays they invoke a rather extensive querying phase at the end (e.g., fact-checking) while for other essays, the same authors submitted all queries way ahead of essay completion. Search engines could leverage that knowledge to support writers with pre-processed search results aimed at supporting fact-checking close to essay completion—a "slow search" way of exploiting the "idle" times when the author is writing. Another idea would be to offer very diverse search results for the first queries of a boil-down writer while build-up writers probably benefit more from more similar search results for their individual queries.

Finally, we also examine different working phases during essay writing and how they influence querying. Although the number of queries alone is a bad predictor for task completion, a general trend is that the number of queries, clicks and pastes decrease over time and the number of writing interactions increases—late fact-checking as an exception.

As for future work, we envision support tools for writers involved in exploratory search tasks as an interesting direction. First steps could be tried along our above described findings (treating build-up and boil-down writers differently, pre-computed fact-checking results for the final phase, etc.). Testing such tools probably then again requires similar studies but our embedding in the TREC environment (topics and the ClueWeb09 corpus) should make comparisons to our findings rather straightforward.

Another interesting direction would be to further analyze the documents the authors used as references and how they were found in the process and to what extent they inspired the final text. Potentially, this kind of analyses can result in better usefulness prediction approaches. Interestingly, on most longer click trails that contain one document the author used as a reference, the authors did find other reference documents, too. These documents not contained in the initial results are good candidates for shortcuts that may be provided to other search engine users in similar situations. A future investigation could also more deeply examine why queriers select their reference documents more carefully than clickers seem to do. It would also be interesting to test the conjecture that authors are not really able to improve their precision in terms of needed queries and clicks to find further reference documents in the course of writing even though having acquired topic knowledge along the process.

References

[1] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *SIGIR 2006*, pp. 3–10.

[2] E. Agichtein, R. White, S. Dumais, and P. Bennet. Search, interrupted: Understanding and predicting search task continuation. In *SIGIR 2012*, pp. 315–324.

[3] J. Arguello. Predicting search task difficulty. In *ECIR 2014*, pp. 88–99.

[4] R. Baraglia, F. Cacheda, V. Carneiro, D. Fernández, V. Formoso, R. Perego, and F. Silvestri. Search shortcuts: a new approach to the recommendation of queries. In *RecSys 2009*, pp. 77–84.

[5] P. Bennett, R. White, W. Chu, S. Dumais, P. Bailey, F. Borisyuk, and X. Cui. Modeling the impact of short- and long-term behavior on search personalization. In *SIGIR 2012*, pp. 185–194.

[6] J. Callan, M. Hoy, C. Yoo, and L. Zhao. The ClueWeb09.

[7] H. Cao, D. Jiang, J. Pei, E. Chen, and H. Li. Towards context-aware search by learning a very large variable length Hidden Markov Model from search logs. In *WWW 2009*, pp. 191–200.

[8] M. De Leeuw and E. De Leeuw. *Read better, read faster.* Penguin Books, 1965.

[9] B. Dervin. From the mind's eye of the user: The sense-making qualitative-quantitative methodology. In *Qual. Res. in Inf. Managm.*, 9:61–84, 1992.

[10] E. Efthimiadis. Query expansion. *ARIST*, 31:121–187, 1996.

[11] Y. Egusa, H. Saito, M. Takaku, H. Terai, M. Miwa, and N. Kando. Using a concept map to evaluate exploratory search. In *IIiX 2010,* , pp. 175–184.

[12] M. Hagen, J. Gomoll, A. Beyer, and B. Stein. From search session detection to search mission detection. In *OAIR 2013*, pp. 85–92.

[13] B. Jansen, A. Spink, and I. Taksa. Research and methodological foundations of transaction log analysis. In *Handbook of Res. on Web Log Analysis*, pp. 1–16, 2008.

[14] K. Järvelin and P. Ingwersen. Information seeking research needs extension towards tasks and technology. *Information Research*, 10(1), 2004.

[15] M. Kurth. The limits and limitations of transaction log analysis. *Library Hi Tech*, 11(2):98–104, 1993.

[16] J. Liu and N. Belkin. Searching vs. writing: Factors affecting information use task performance. *ASIST*, 49 (1):1–10, 2012.

[17] G. Marchionini. Exploratory search: From finding to understanding. *CACM*, 49(4):41–46, 2006.

[18] H. O'Brien and E. Toms. What is user engagement? A conceptual framework for defining user engagement with technology. *JASIST*, 59(6):938–955, 2008.

[19] H. O'Brien and E. Toms. The development and evaluation of a survey to measure user engagement. *JASIST*, 61(1):50–69, 2010.

[20] U. Ozertem, O. Chapelle, P. Donmez, and E. Velipasaoglu. Learning to suggest: A machine learning framework for ranking query suggestions. In *SIGIR 2012*, pp. 25–34.

[21] M. Potthast, M. Hagen, B. Stein, J. Graßegger, M. Michel, M. Tippmann, and C. Welsch. ChatNoir: A search engine for the ClueWeb09 corpus. In *SIGIR 2012*, p. 1004.

[22] M. Potthast, M. Hagen, M. Völske, and B. Stein. Crowdsourcing interaction logs to understand text reuse from the web. In *ACL 2013*, pp. 1212–1221.

[23] M. Potthast, M. Hagen, M. Völske, and B. Stein. Exploratory search missions for TREC topics. In *EuroHCIR 2013*, pp. 11–14.

[24] Y. Qu and G. Furnas. Model-driven formative evaluation of exploratory search: A study under a sensemaking framework. *IPM*, 44(2):534–555, 2008.

[25] S. Robertson, H Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *CIKM 2004*, pp. 42–49.

[26] D. Russell, M. Stefik, P. Pirolli, and S. Card. The cost structure of sensemaking. In *CHI 1993*, pp. 269–276.

[27] E. Sormunen, J. Heinström, L. Romu, and R. Turunen. A method for the analysis of information use in source-based writing. *Information Research*, 17(4): paper 535, 2012.

[28] B. Stein, T. Gollub, and D. Hoppe. Beyond precision@10: Clustering the long tail of web search results. In *CIKM 2011*, pp. 2141–2144.

[29] J. Teevan, K. Collins-Thompson, R. White, and S. Dumais. Slow search. In *CACM*, 57(8):36–38, 2014.

[30] P. Vakkari. Exploratory searching as conceptual exploration. In *HCIR 2010*, pp. 24–27.

[31] P. Vakkari and S. Huuskonen. Search effort degrades search output but improves task outcome. *JASIST*, 63 (4):657–670, 2012.

[32] R. White and G. Marchionini. Examining the effectiveness of real-time query expansion. *IPM*, 43(3): 685–704, 2007.

[33] R. White and R. Roth. *Exploratory search: Beyond the query-response paradigm.* Morgan & Claypool, 2009.

[34] R. White, B. Kules, S. Drucker, and M. Schraefel. Introduction. In *CACM*, 49(4):36–39, 2006.

[35] R. White, G. Marchionini, and G. Muresan. Evaluating exploratory search systems. *IPM*, 44(2):433–436, 2008.

[36] J. Xu and B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM TOIS*, 18(1):79–112, 2000.