Declaration of Authorship

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Weimar, September 5th, 2014

.....
Jakob Gomoll

Abstract

This work examines and characterizes users' interacting behavior in exploratory search tasks, i. e., investigative tasks on a domain that a user is not entirely familiar with. For the purpose of characterization, a variety of analyses is conducted on two corpora, which are provided by the Webis group. The first one, the Webis-Query-Log-12, contains detailed information about all interactions with a search engine that users did while working on such an exploratory search task. The second corpus, the Webis-TRC-12, consists of the essays that were produced by the same users with respect to their exploratory tasks. So, it is not only known what a user has searched for but also what the actual user's task was and how much the submitted queries and visited documents helped the user to complete this task.

The contributions of this thesis are twofold: First, the Webis-Query-Log-12 is depicted in detail as it has been largely reorganized since its original release. Second, both aforementioned corpora are examined hand in hand to find patterns in users' interacting behavior. In this context, it is shown that users often have a guiding query, which attends the entire search process. Further, the existence of two elementary searching strategies is revealed by this work: one group of search engine users focuses on formulating rather exact queries to find a particular needed information, whereas the other group submits only few queries and finds the desired information by clicking from document to document. Yet, the analyses also show that the applied searching strategy seems to have neither a positive nor a negative impact on the quality of users' work. Beyond this, the work at hand picks up again the two text reuse patterns *build-up* and *boil-down* identified by Potthast et al. [PHVS13a], refines them by introducing an unambiguous, computable value for essay growth, and further extends it by the variable paste regularity. This revised classification scheme reveals that users typically pursue an individual working style, which distinguishes each author from another.

Contents

1 Introduction

The aim of this work is to characterize users' interacting behavior in exploratory Web search tasks, in which users search for information on an unfamiliar domain [WR09]. As a base for our investigations, two corpora are analyzed side by side: One contains the revision history of 150 user-written essays, and the other contains all interactions accompanying the writing processes, i. e., submitting queries to a search engine, clicking on shown results, and navigating from site to site.

In Information Retrieval, exploratory search is the process of investigating a topic that is, at least in parts, unfamiliar to a user. It typically starts with only a rough idea about the target domain, and the user may not be equipped with topic-specific vocabulary nor knowledge about the information space structure [WKD+06]. By formulating rather general queries at the beginning, the user explores this space and acquires more knowledge about it. Learned keywords and concepts can then be used in the next iteration of the search process, which leads to more sophisticated queries and a more deliberate selection of search results to visit and learn from. As a result of the continuously growing knowledge, goals in exploratory search tasks often emerge and change several times during the whole process.

However, although the learning process described above seems quite natural to us, it has hardly been proven until today. The work at hand will present and examine the Webis-Query-Log-12 corpus [PHVS13b], which contains fine-grained interaction logs of users performing exploratory search tasks. The Webis-Query-Log-12 differs from conventional corpora as it not only contains plain search interactions but also information about the current state of task completion: Users were asked to write an essay about one of 150 well-defined TREC topics from the recent years; and while they did so, each iteration of the essay's formation process was recorded and aggregated into the text reuse corpus by the Webis group, referred to as Webis-TRC-12 [PHVS13a]. Thus, by linking the available information from both corpora, it is possible to examine users' interacting behavior at any stage of their exploration process.

This exhaustive amount of information allows us to gain deep insights into how people are using search engines in a setting that corresponds to a realistic task as closely as possible. To the best of our knowledge, only few similar datasets exist, but these either lack diversity of topics (cf. only two different topics in [QF08]) or relevance for every-day

life (cf. studies in a medical context in [VH12]). Accordingly, as of the time writing this thesis, the Webis-Query-Log-12 and the Webis-TRC-12 form the largest and probably most realistic, publicly available[1] data source for applying research studies on exploratory search behavior.

Working on these comprehensive data sets certainly raises a variety of questions. However, the main research question of this thesis is as follows: Which patterns can be found in the log data that characterize users who are working on an exploratory search task? In this regard, it is also discussed if any of these patterns and characteristics could be automatically detected during an ongoing search process, and whether search engines could make use of them to better support users in achieving their current goal.

In order to get some better understanding of the challenges that are faced in this work, Chapter 2 briefly presents an overview of related work. Further, we discuss prior approaches for analyzing users' interacting behavior and how these relate to this thesis. Subsequently, Chapter 3 presents the aforementioned corpora along with some key figures. However, the examination of the Webis-Query-Log-12 is of superior interest as it has been largely reorganized since its initial release. Chapter 4 constitutes the main part of this work and describes the analyses that were conducted in the course of this thesis. It examines the structure of queries, identifies useful documents, reveals and distinguishes between two elementary searching strategies, it measures the user engagement, clusters author types into their respective writing strategies, and finally analyzes how different working phases relate to each other. Additionally, we discuss the relevance of these findings for today's Web search engines. Chapter 5 sums up the key outcomes of this work and shows several directions for further research.

[1] `http://webis.de/research/corpora/`, Last accessed: August 30th, 2014

2 Related Work

Nowadays, almost every need for information can be met by the Web, and people across all ages are used to work with Web search engines. While searching algorithms have improved a lot over the last years, researchers are still trying to get a better understanding of cognitive processes that underly a Web search. Perhaps, one of the most important findings is the differentiation between two fundamental types of search tasks: closed-ended and open-ended tasks, speaking in the vocabulary of Marchionini [Mar06]. The former is a search for a particular fact, e. g., a famous person's date of birth, whereas the latter does not necessarily have a clear outcome, e. g., finding the best hotel for an upcoming trip. Open-ended tasks often do not lead to only one correct answer but they help constituting a mental model [Vak10] of a topic and/or forming an opinion. Speaking about open-ended tasks, the term *exploratory search* is often used as a synonym.

2.1 Exploratory Search

According to White and Roth, exploratory search describes an *information-seeking process* taking place in an "open-ended, persistent and multi-faceted" problem context [WR09]. They explicitly refer to information-seeking instead of Information Retrieval (IR), because they argue that in IR the presence of a searched item is certain, whereas in information-seeking it is unknown whether there even exists a document that can fulfill the information need. Therefore, they further describe exploratory search as an "opportunistic, iterative, and multi-tactical" process [WR09], in which users explore the information space as extensively as wanted or needed. However, describing this process as iterative does not necessarily mean that iteration steps cannot occur in closed-ended search tasks. Yet, in these tasks, query iterations may narrow down to one specific target document, whereas iteration steps in exploratory search usually serve to explore different aspects of the information space (cf. Figure 2.1).

For better imagination, Figure 2.1 provides an illustration of a possible information space exploration during an exploratory search (right side of the figure). Starting with a "tentative query, [users] navigate proximal to relevant documents," where they start to "explore the environment to better understand how to exploit it" [WR09]. While examining search results, users passively obtain clues for their next steps [WMM08].

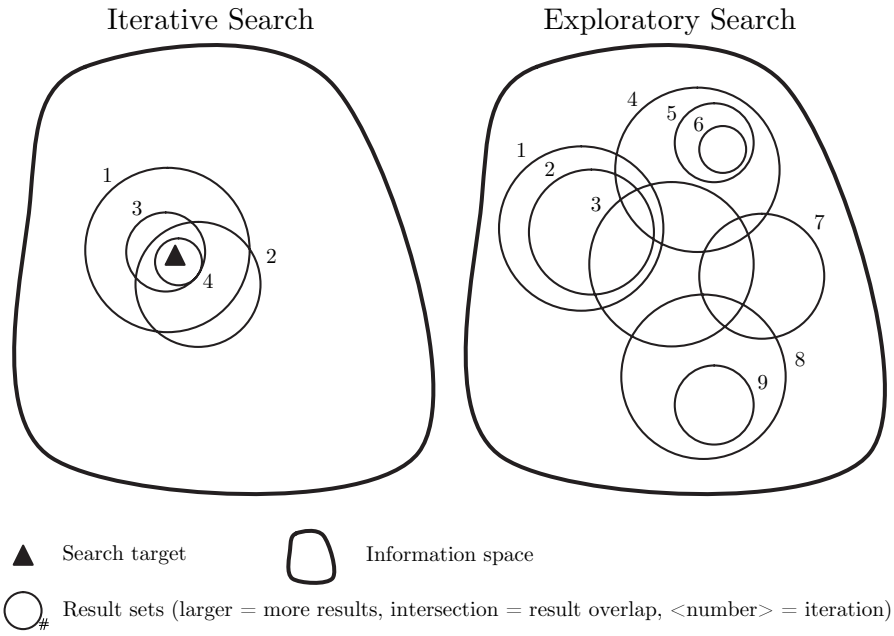


Figure 2.1: Iterative search strategy in information Retrieval versus exploratory search in Information-seeking (figure from [WR09]). Note that iterative search aims at finding a particular document while exploratory search focuses on exploring the information space as extensively as possible.

In Figure 2.1, the total area covered by the circles represents the user's knowledge within the surrounding information space. Note that this area typically is much larger in exploratory than in iterative, closed-ended search, where only a particular document was of interest.

The big challenge in the field of exploratory search is to design new retrieval models that support users in their exploratory tasks. Common search engines are typically tuned towards precision, i. e., they try to exclude possibly irrelevant items [Mar06], and thereby restrain the chance of broadening the user's horizon. Instead, the nature of exploratory search is rather related to recall, i. e., finding many possibly relevant items [Mar06] and revealing multiple aspects of the target domain, with the chance, however, of retrieving some irrelevant items. Therefore, new systems should provide rapid query refinement, especially in the early phase of a search (as in [WM07]), support facets, for example by clustering search results (as in [SGH11]), and leverage the search context, e. g., by query expansion with the help of keywords from top-ranked documents [XC00].

2.2 Analyzing Search Behavior

Learning about users' search behaviors from interaction logs is a valuable starting point for designing systems that aim to support users in their search tasks. Although researchers have been working on this topic for many years, search log analysis still remains a big challenge in the field of Information Retrieval. As Kurth argues, all measures that can be derived from plain user interactions cannot explain the intention behind these interactions [Kur93]. For example, the fraction of unique query terms to all query terms does not necessarily correlate with users' engagement. However, asking users in a research study for their intention behind each interaction may provide remarkable insights, but would only allow to design a supporting system in an obtrusive environment. Since users prefer simplicity and do not like to declare their motivation for each query or click, such a system is simply not appropriate for use in everyday life.

Therefore, researchers have to rely on measures that can be derived from plain interactions as these are the only utilizable clues from users dealing with a usual IR system. Typical measures that can be found in the literature comprise the number of queries issued by a user, the average number of terms per query, as well as the average number of clicks per query, or the time between query and first click [Arg14]. Furthermore, attributes from a more global context, like the number of physical sessions spanned by this task, are often used in log analyses. Machine learning algorithms then combine a wide range of such measures in order to train classifiers for particular tasks the researchers are interested in (e. g., [ABDR06, CJP+09, BWC+12, OCDV12]).

Agichtein et al., for example, were interested in predicting if a user is likely to resume a suspended session within the next few days [AWDB12]. Notable about this work is that the authors used the Open Directory Project[1] in order to assign queries to topical categories. Once the dominant topic of the majority of queries was determined, they were able to automatically decide for each query whether it is related to the task or not. Thus, a variety of statements about users' engagement and focus, e. g., the ratio of on-task queries to all queries in a session, could be used as additional features for classifier training. In the end, user engagement and focus turned out to be the most discriminating feature group for the aimed prediction [AWDB12].

However, many of these approaches deal with the typical dilemma of machine learning that outcomes are often unintuitive. Although researchers precisely describe what they did to train their classifier and which features they used, they lack a conceivable explanation for *why* their classifier works. In this regard, users' behavior is not actually characterized what we aim for. Further, the trained classifiers heavily rely on the used interaction log, which often is neither publicly available nor comparable to other data sets.

[1] `http://www.dmoz.org/`, Last accessed: August 30th, 2014

2.3 Analyzing Search Behavior in Exploratory Search

While analyses based on machine learning algorithms have been applied for more than two decades, the more recent years brought a subtle interest in analyzing users' interacting behavior in exploratory search.

To our knowledge, the first research study on this topic was conducted in 2008 by Qu and Furnas [QF08]. Based on the sense-making model [Der92, RSPC93], they designed a study that should reveal the relation between information seeking and construction of a mental representation. For that purpose, they recorded not only users' interactions with the search system but also asked the 30 participants to prepare an outline for a 1-hour talk; thus, the users provided an external representation of their understanding of the given topic. Interestingly, Qu and Furnas found out that this final talk outline strongly correlated with people's bookmark structure, i. e., the folders in which bookmarked pages were organized [QF08].

Further, human judges rated the topical similarity between consecutive queries and assigned each query to one of the bookmark folders. On a timeline, Qu and Furnas visualized these information, i. e., made visible when which query occured, which folder it referred to, and which web page was bookmarked in this context (cf. Figure 2.2). They claim that the visualizations for all of their 30 subjects clearly reveal the influence of recently constructed representations on the upcoming search process. Moreover, they found that 14 out of 30 participants used their folder structure as a roadmap for their upcoming search. In some cases, people even searched the Web for new structure ideas, i. e., they searched for possibly existing outlines [QF08]. The authors conclude that new search systems should consider user's representation construction process (for instance by automatically analyzing the bookmark structure) and support this process.

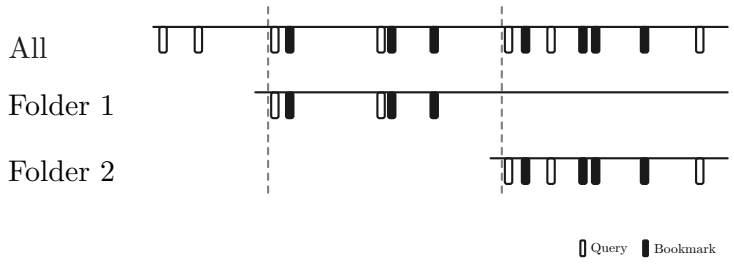


Figure 2.2: Log visualization schema proposed by Qu and Furnas [QF08]. The figure shows an imaginable querying history along with some pages bookmarked in some specific folder. The vertical lines denote a topical change, i. e., the previous query is topically different from the following one.

A quite similar approach was pursued by Egusa et al., who asked users to produce a concept map of their understanding of a given topic [EST+10]. A concept map here is a graph consisting of named entities and labeled connections between them. Participants were asked to provide such a concept map before and after the search on the given topic. By analyzing the differences between both maps, Egusa et al. tried to develop a new evaluation measure for exploratory search tasks that can assess the task performance. Such a new measure is needed because common search engines are tuned towards precision (cf. Section 2.1), and consequently, traditional measures from IR only assess the precision of a result list but not the *benefit* a user has from the shown results [WR09]. In this regard, Vakkari differentiates between *search engine output*, which basically is 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 [Vak10]. Note that performance of task outcome can be independent from the quality of the search engine output, i. e., a high precision does not necessarily lead to good task performance.

However, Egusa et al. did their experiments with 35 undergraduate students on only two different but truly open-ended topics: Politics and Media. The task was to find multiple opinions about these two topics and compare each of the found view points [EST+10]. Because of the topical complexity, the researchers could not actually concentrate on qualitative aspects of the pre- and post-search concept maps but rather on quantitative descriptions. Thus, they analyzed the number of kept, discarded and new nodes (or links and labels, respectively) and found out, for example, that nearly as many nodes, links and labels were discarded as new ones were created. This clearly shows that people not only gather new input through exploratory search but also adapt their already existing knowledge structures [EST+10]. However, they conclude that applying descriptive statistics on concept maps cannot serve as a measure for performance of an exploratory search system. They argue that one has to conduct more qualitative analyses of the described concepts and users' seeking behaviors.

Against this background, Vakkari and Huuskonen designed a study that rather concentrates on the search process, especially the effort that users put into the search [VH12], and how it is interlinked with task outcome (cf. Vakkaris claim from above). Within the scope of a semester'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 then examined in respect of applied search tactics (e. g., 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) [VH12]. The essays' evaluation scores, which were given by the teachers, were used as a performance measure for task outcome. By applying a so-called path analysis, a specialized regression model [PK82], Vakkari and Huuskonen were able to show several interesting linkages between search process, output and outcome variables. As major outcome they constitute the negative correlations between diversity of queries, search engine's precision, and essays' scores.

That means, the broader the queries were formulated, the lower the system's precision was, yet the higher the essay scores were. A very similar correlation was observed for search effort: The more sessions a user needed to write the essay, the lower was the system's overall precision because of the larger result set, but at last it resulted in a higher quality of the essay [VH12].

However, Vakkari and Huuskonen themselves admit that there still are a lot of unknown conditions that influence the task outcome, which should be studied in further research. They conclude that more studies are needed that focus on how the retrieved information is used, how it contributes to user's understanding and how changes in this understanding are reflected in the upcoming retrieval process [VH12].

Finally, Potthast et al. constructed and published a corpus that should facilitate such researches [PHVS13b]. The Webis-Query-Log-12 contains detailed information about users' queries, the use of facets, search results that were clicked, and supplementary web pages users visited in a click trail. In their publication, Potthast et al. describe the log, its construction and present first quantitative descriptions as well as some basic qualitative investigations. Using this log in combination with the Webis-TRC-12 [PHVS13a] allows us to conduct manifold analyses, which overcome several issues of the previously presented work:

1. **Nature of topics:** In the studies conducted by Qu and Furnas [QF08] as well as Egusa et al. [EST+10], participants searched on only two different topics. Vakkari and Huuskonen [VH12] provided at least eleven different topics, which were elaborated by a highly specialized group of people as the topics were from a medical context. In this work, we examine whether their findings also hold for more general topics.
2. **Users' mental representations:** In contrast to the presented works by Qu and Furnas [QF08] or Egusa et al. [EST+10], we do not rely on some external representation (cf. talk outline/bookmark structure, or concept map, respectively) but have direct access to the current status of the essays that users are actually writing while working on a topic.
3. **Use of available data:** The type of data that Vakkari and Huuskonen [VH12] used for their analyses is quite analogous to ours, i. e., they had access to a search interaction log and the produced essay. Yet, they relied on meta-data (the teachers' evaluation scores) instead of the essays themselves. In the work at hand, we substantiate our conclusions by directly using the produced contents.
4. **Environmental conditions:** Qu and Furnas [QF08] required their participants to use a customized search system and thereby produced an obtrusive environment. In our case, people were allowed to work anywhere they like with the browser they normally use, thus imitating a more realistic scenario.

3 Webis-Query-Log-12 and Webis-TRC-12

As already stated, the work at hand examines two logs, namely the Webis-Query-Log-12 and the Webis-TRC-12. In the following sections, an overview of both logs and their relation will be given first, and then each of the logs will be depicted in detail along with descriptive key figures.

3.1 Log Overview

Both logs were compiled in the context of the PAN competition [PGH+12], in which computer scientists are encouraged to implement plagiarism detection algorithms. From 2009 to 2011, the test documents for this competition were automatically generated and obfuscated, i. e., text passages from random documents were inserted into host documents and restructured by randomly shuffling and replacing words [PHVS13a]. Although detecting the plagiarized passages in the resulting documents may have been quite difficult for algorithms, it was an easy task for humans.

So, the corpora from these years were far from reality, and a new ground truth was desired. Since no other sufficient data set was available back in 2012 [PHVS13a], the Webis group[1] spent more than 20,000 $ to hire (mostly professional) writers to produce a decent corpus. The hired authors were asked to write essays, in which they were encouraged to generously reuse text passages from third-party authors. Yet, each author should rephrase the plagiarized passages so that even humans would have trouble identifying them. For the purpose of traceability, the authors themselves designated each plagiarized passage to its source document.

In total, 297 individual essays on 150 different topics were collected. Each topic was treated in two different scenarios: in one scenario the sources to plagiarize from were pre-selected and the usage of any additional source was prohibited, whereas in the other scenario authors had to retrieve source documents themselves using the ChatNoir search engine [PHS+12]. All interactions from the second scenario were logged and compiled into the Webis-Query-Log-12. The iteration steps of the essays, i. e., the single revisions, were compiled into the Webis text reuse corpus 2012, referred to as Webis-TRC-12.

[1] `http://webis.de/`, Last accessed: August 30th, 2014

3.2 Webis-Query-Log-12

The Webis-Query-Log-12 consists of 150 XML files containing all interactions that have been recorded during the work on one topic, i. e., one file stands for one topic. Basically, there are three different interaction types:

1. **Query** – A query consists of the search terms issued by the user along with used facets and shown results. For each result, information about the document's rank in the result list, the ClueWeb document ID, the real URL, the title and the snippet generated for that particular query are available.
2. **Click** – A click is characterized by the ClueWeb ID, the real URL and title of the visited document. Further, the type of its origin is denoted, so a click can be classified as either a result click, a so-called trail click, or a bookmark click. A result click occurs when a user visits one of the documents presented on the search engine result page. From that document, the user might follow several hyperlinks to subpages or other documents, what we call trail click. The remaining clicks, for which the origin could not be ascertained from the raw log data, are classified as a bookmark click when the document has been visited in the past.
3. **Text-writing** – The text-writing interactions are not actual interactions with the search engine but serve as a cross reference to the Webis-TRC-12 corpus. It consists of the start as well as the end revision of the text-writing interaction and their respective timestamps. Note that queries and clicks cannot occur within such a text-writing block.

Further, each of the above interaction types has a timestamp and an anonymized IP address, which gives a clue about different workstations that users worked with. See Listing 3.1 for a sample excerpt of the Webis-Query-Log-12.

All in all, the Webis-Query-Log-12 contains 13,609 queries, 16,698 clicks and 6,123 text-writing interactions by 12 different users. All interactions took place in the period from mid of June 2012 to end of November in the same year, spanning 166 days (about five and a half months) in total. The longest time period that has been spanned by one topic is 56 days, yet the author actually worked on only 12 days on this topic and paused work on the other 44 days. The majority of authors worked 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 being absent from the computer after 15 minutes of inactivity – the log contains 2,797 physical sessions, resulting in an average of 18.6 sessions per topic or 3.4 sessions per working day, respectively. The shortest sessions only

```
<?xml version="1.0" encoding="UTF-8"?>
<interaction_log>
  <topic id="t110" user="u002">
    <title>Brazil Topography</title>
    <description>
      Write about Brazil's topographical features and its geopolitical
      situation. Is Brazil a safe destination for vacations?
    </description>
    <interaction id="t110-i1">
      <timestamp>2012-08-28T05:55:44</timestamp>
      <ip>127.0.0.34</ip>
      <query>
        <string>brazil geography</string>
        <facets>
          <desired_num_results>10</desired_num_results>
          <long_text>false</long_text>
          <read_level>None</read_level>
        </facets>
        <shown_results>
          <result id="t110-i1-r1">
            <rank>1</rank>
            <doc_id>1300408759</doc_id>
            <warc_trec_id>clueweb09-en0130-04-08759</warc_trec_id>
            <url>http://www.brazil-travel-northeast.com/geography-of-brazil.html</url>
            <doc_title>Geography Of Brazil</doc_title>
            <snippet>
              Geography Of Brazil > Geography of Brazil northeast > Historic Center Salvador
              All along the coast, you will find very many historic and most beautifully
              preserved colonial towns and cities ...
            </snippet>
          </result>
          <!-- ... snap ... more results ... -->
        </shown_results>
      </query>
    </interaction>
    <interaction id="t110-i2">
      <timestamp>2012-08-28T05:55:49</timestamp>
      <ip>127.0.0.34</ip>
      <click>
        <doc_id>1300408759</doc_id>
        <warc_trec_id>clueweb09-en0130-04-08759</warc_trec_id>
        <url>http://www.brazil-travel-northeast.com/geography-of-brazil.html</url>
        <doc_title>Geography Of Brazil</doc_title>
        <reference type="result-click">t110-i1-r1</reference>
      </click>
    </interaction>
    <interaction id="t110-i3">
    <timestamp>2012-08-28T05:58:38</timestamp>
      <ip>127.0.0.34</ip>
      <text-writing>
        <start-time>2012-08-28T05:58:38</start-time>
        <start-revision>3</start-revision>
        <end-time>2012-08-28T06:48:14</end-time>
        <end-revision>161</end-revision>
        <duration>PT0H49M36S</duration>
      </text-writing>
    </interaction>
    <!-- ... snap ... more interactions ... -->
  </topic>
</interaction_log>
```

Listing 3.1: Sample excerpt of Webis-Query-Log-12

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 4.1.

Table 3.1 summarizes these and other key figures from the Webis-Query-Log-12. It reveals several interesting and possibly unexpected findings. Perhaps, the most surprising finding is the ratio between unique queries to all submitted queries, which is about one quarter (3,538 to 13,609). Indeed, almost every query has an identical follow-up query requesting more results (6,874 queries with 10 results and 6,727 follow-up queries requesting 100 results). Although this explains why the number of unique queries should not exceed 50 % of all submitted queries, an explanation why it is only about 25 % is still missing. Part of this explanation is provided in Section 4.2, which will show that many users quite often submit identical queries in a row.

It is also remarkable that in more than half of the physical sessions no query at all was submitted to the search engine. In fact, the third quartile of queries per session lies at only 4, meaning that 75 % or 2,097 of all sessions contain 4 or even fewer queries. Conversely, this means that the majority of submitted queries (12,094 out of 13,609) was issued in only 700 sessions. The number of clicks per session shows a very similar distribution of values: here, 14,421 of 16,698 clicks take place in only 700 sessions. This indicates that text-writing interactions form the largest part of most physical sessions.

Table 3.1: Key figures describing the Webis-Query-Log-12.

	Min	Q1	Median	Average	Q3	Max	$\sum$
Meta-data							
Spanned days per topic	1.0	4.0	6.0	8.6	9.0	56.0	-*
Spanned working days per topic	1.0	4.0	5.0	5.5	7.0	17.0	-*
Working hours per topic	1.8	5.2	7.5	7.9	9.8	23.0	1,191
Physical sessions per topic	2.0	11.5	16.0	18.6	23.0	55.0	2,797
Queries							
Queries per topic	4.0	40.0	68.0	90.7	117.0	612.0	13,609
Unique queries per topic	1.0	12.0	20.0	23.6	31.5	121.0	3,538
Queries per session	0.0	0.0	0.0	4.9	4.0	231.0	13,609**
Clicks							
Clicks per topic	12.0	55.0	87.0	111.3	144.5	431.0	16,698
Clicks on individual documents per topic	8.0	44.5	67.0	74.5	101.0	259.0	11,181
Result clicks per topic	5.0	30.5	49.0	58.5	75.5	280.0	8,779
Trail clicks per topic	0.0	13.5	33.0	52.8	73.0	332.0	7,919
Clicks per session	0.0	0.0	1.0	6.0	6.0	164.0	16,698**
Clicks per query	0.0	0.0	0.0	0.2	0.0	76.0	8,779**
Reading time per click (minutes)	0.0	0.1	0.4	0.7	0.8	15.0	11,236
Text-writing interactions							
Text-writing interactions per topic	11.0	28.0	42.0	46.3	59.5	178.0	6,943
Time per text-writing interaction (minutes)	0.0	0.5	2.2	7.4	8.9	145.2	51,126

* Summing up these values might lead to misinterpretation of the data.

** Note that these values are naturally equal to one of the priorly listed.

A closer inspection of the click statistics reveals two other interesting aspects. First, the number of result and trail clicks is quite balanced, which indicates that people genuinely followed exploratory search strategies and not just entered look-up queries. Second, users spent less time reading the clicked documents than we expected. 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 spending much time for reading them. For example, only 661 documents, which is about 4 % of all clicks, were viewed for two and a half minutes or longer. Yet, Section 4.3 will show that those documents are not necessarily of higher relevance for authors than documents with less reading time.

Last but not least, it is remarkable which extensive amount of workload and time the authors spent on their tasks. Resulting in a total of 1,191 working hours (49.6 days), users spent about 152 hours (6.3 days) for querying and examining the result pages, approximately 187 hours (7.8 days) for reading the retrieved documents, and a substantial quantity of 852 hours (35.5 days) for essay writing. However, note that only the essay writing time is exactly known and the durations for query and click interactions tend to be overestimated as they may include short breaks. Those breaks occur, for example, when users submit a query, then leave their working place to get some coffee, and return to it after a couple of minutes. As long as this phase of inactivity lasts less than 15 minutes (the physical session cut-off time determined in Section 4.1), it is not clear whether the user really studied the result list for such a long time or just enjoyed a short break. A superficial analysis showed that those breaks could form up to 122 hours of the total working time, which corresponds to an average of 49 minutes per topic. This, in turn, seems appropriate with regard to an average working time of 7.9 hours per topic.

While the key figures described above may convey a rough impression of the overall data, Figure 3.1 shows the distribution of some of these values along the different users. It can be seen that user 002 was the most dedicated one, especially in terms of time spent per topic. With a median working time of 13 hours per topic, this author outperforms almost all others (the only exception is user 021 who worked on one single topic for 14 hours). Further, a coherence between spanned working days, working hours and physical sessions can be deduced from the plot: the more different working days were encompassed for topic treatment, the higher was the total amount of working time, and the more physical sessions took place. This is actually not surprising but confirms our expectations of typical user behavior.

In contrast to this, more queries do not necessarily entail more clicks as can be observed for users 017 and 021, for example. Although user 017 submitted considerably more queries per topic than user 021 (compare a median value of 106 queries to a maximum value of 96), the number of clicks shows the exact opposite (here, user 017 did 143 clicks at most, whereas the median value for user 021 is 236). One implication could be that different types of searchers exist: clickers and finders. Clickers tend to submit an initial

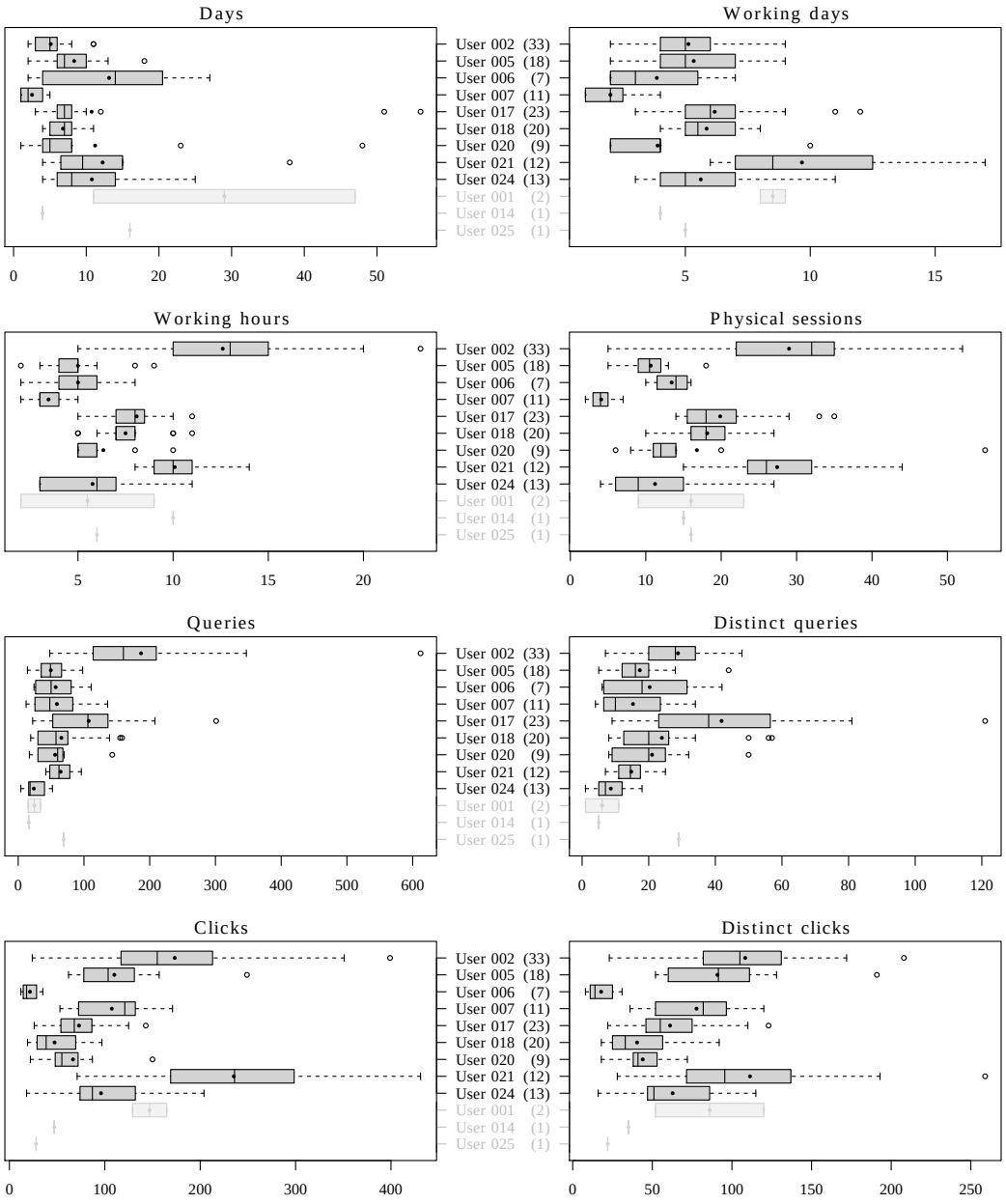


Figure 3.1: User-wise distribution of key figures. The number in brackets behind each author gives the number of treated topics, and the filled circles in the plots represent the average value for the respective measure. Users 001, 014 and 025 are grayed out as they worked on few topics only.

query and then follow a click trail along various documents until they find the desired information, whereas finders refine their queries in an iterative process to narrow down to the searched information. Section 4.4 will reveal that these different strategies really can be found along the users.

Since its initial release in 2013 [PHVS13b], several minor and major adaptations have been applied to the Webis-Query-Log-12. The most obvious change is the introduction of a well-defined format. Whereas search interactions in the original log were typically accessed via database requests, they are now organized in an XML structure, which has several advantages:

1. The log is easily portable, i. e., it is not necessary anymore to export and import database dumps between different machines, which often entails problems like conflicts in users' rights management, etc.
2. The log is human readable. One of the greatest benefits of an XML document compared to a database is that all available information are accessible at one place. For example, clicks now contain a direct reference to their origin and inconvenient JOIN operations between different tables are no longer required.
3. The log can be easily parsed by many standard programming languages. On top of that, it seems more convenient to work with the data in memory since temporal customizations in the log data may be applied without modifying the original data.
4. It is guaranteed that all data are consistent to a carefully designed XML schema. In the initial version of the Webis-Query-Log-12 many entries were incomplete or used different notations for equal concepts (especially IDs).

Moreover, a couple of incorrect information was detected in the original data set, which were overhauled in the current version. In most cases, those misinformation emerged from unhandled logging errors and comprise duplicate queries, missing log data (in some cases even longer sequences of queries and clicks), false temporal order of interactions, or test queries and clicks originating from the Webis staff when supporting the authors. Table 3.2 shows how many queries and clicks were contained in the original log, how many of them were removed after being identified as duplicates or other erroneously recorded data, and how many missed data was inserted into the new log.

Table 3.2: Comparison between initial and current version of the Webis-Query-Log-12.

	Old Log	Removed Entries	Added Entries	New Log
Queries	13,651	-277	+235	13,609
Clicks	16,739	-303	+262	16,698

3.3 Webis-TRC-12

As already mentioned, the Webis text reuse corpus is the collection of all written essays and was released in 2013 [PHVS13a]. The outstanding property of this log compared to others is that it not only contains the finished essays but also all intermediate states from the beginning up to essay completion. For that purpose, the current state of an essay has been captured as soon as the user has stopped typing for more than 300 milliseconds. This extraordinarily short duration allows analyses on a very fine-grained level.

Table 3.3 shows a few key figures that are of interest for this thesis. First to mention is that average, median, $Q1$ and $Q3$ for essay length are close together not just by coincidence but that the Webis group asked the authors to produce essays with an approximate length of 5,000 words. There are, however, a few essays that are significantly shorter or longer than this intended value. Short essays are the consequence of difficulties in finding useful source documents on the respective topic. For example, one author wrote about the HP Mini 2140 notebook and claimed that its market launch might have overlapped the crawling period of the ClueWeb09 corpus in January and February 2009 [CHYZ09]. Therefore, only announcements of the product could be found, and the author was not able to write a review on this product as demanded by the task description.

Further, the table shows that the number of different documents from which content was plagiarized from is between 11 and 21 for half of the essays. From our natural understanding, this seems an appropriate number of sources for a 5,000-words essay. Only 25 % of the essays contain less than 11 sources with a minimum of only 3 references. In Section 4.5 it is shown that users with many references can be considered slightly more dedicated to the task than users with only few sources.

Beyond this, Table 3.3 contains statistics about the number of *text paste events*. It is important to note that these values are not originally part of the Webis-TRC-12 as these were not recorded by the text editor but only detected by a post-processing step in the course of this thesis. For this purpose, all single essay revisions were cleaned at first, i. e., the questionnaires from the header and footer were removed, before an algorithm has been applied to extract differences between two consecutive essay revisions. If

Table 3.3: Key figures describing the Webis-TRC-12.

	Min	Q1	Median	Average	Q3	Max	$\sum$
Number of revisions	259.0	1,809.5	2,852.0	2,848.8	3,815.0	6,794.0	-*
Essay length (# words)	709.0	4,752.5	4,982.0	5,029.8	5,224.5	13,877.0	-*
Paste events	0.0	13.0	25.0	28.6	39.0	134.0	4,291
Used references	3.0	11.0	16.0	18.4	21.0	69.0	2,761

* Summing up these values might lead to misinterpretation of the data.

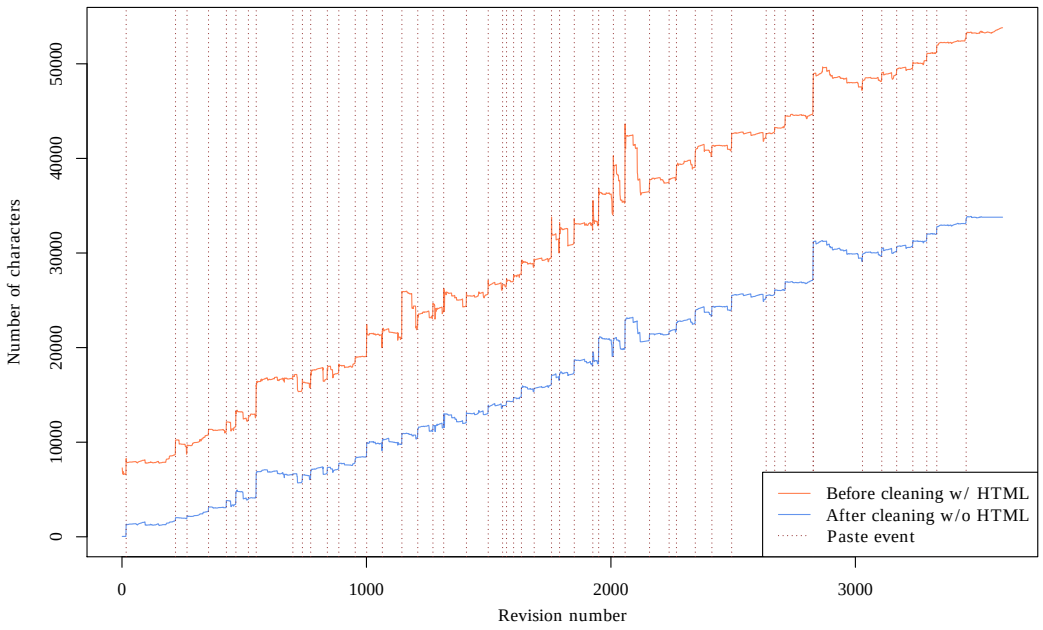


Figure 3.2: Essay completion graph with paste events for topic 29.

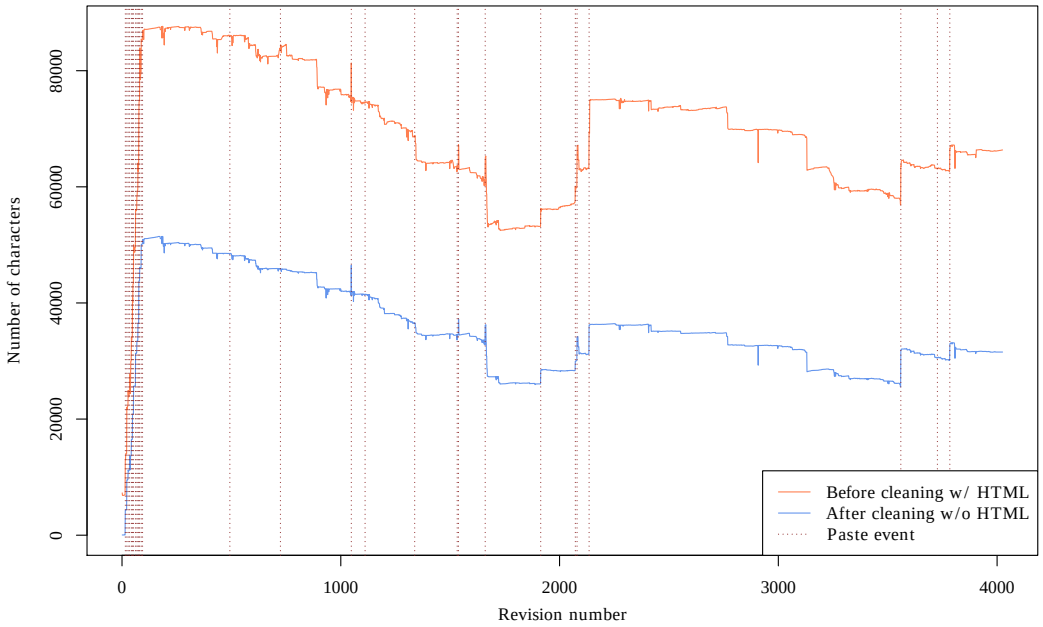


Figure 3.3: Essay completion graph with paste events for topic 50.

more than 100 characters have been added between both revisions, we suppose the added content to have been pasted as it is quite unrealistic that users really typed 100 signs without having made even one short break of 300 milliseconds. For this analysis, paste events have been ignored when the content was removed from the essay before (Ctrl+X, Ctrl+V).

A visual inspection (see Figures 3.2 and 3.3) for each of the 150 topics showed that this detector worked quite accurately. The figures show the logical time on the x-axis and the essay growth on the y-axis. The blue line describes the length of the actual essay content without HTML residues, and the orange line is shown as it sometimes can indicate such a paste event where it is not evident on the blue line (cf., the last paste event before revision 1,000 in Figure 3.3). The vertical, dotted lines represent the detected paste events. These two figures have been chosen as they provide a first impression of different writing strategies followed by different authors, which will be treated in Section 4.6. Here, it is clearly visible that the author of topic 29 pasted content into the essay more or less continuously, whereas paste events in topic 50 show a rather irregular pattern.

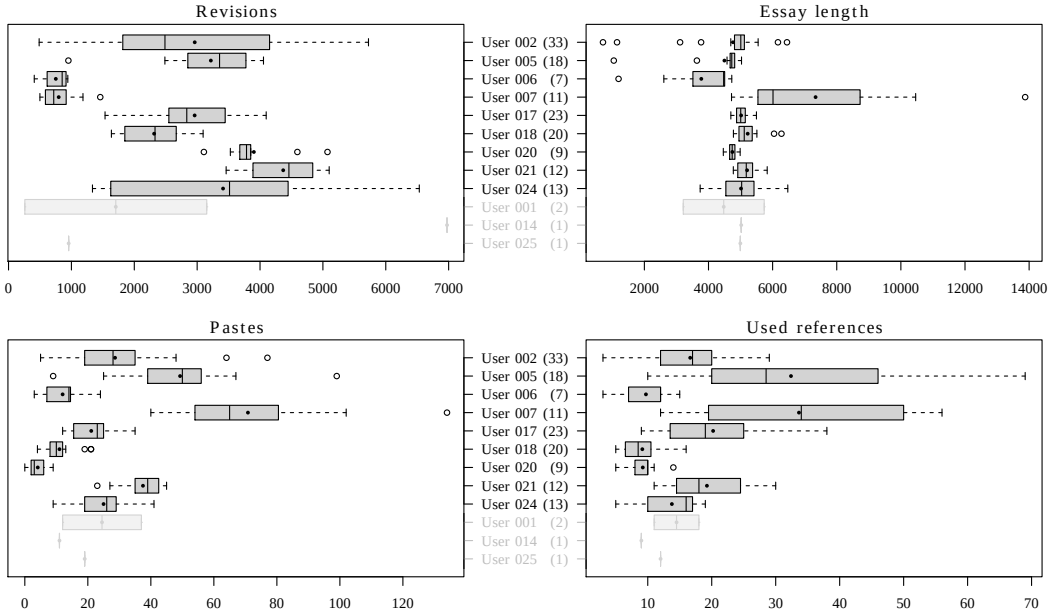


Figure 3.4: User-wise distribution of key figures in Webis-TRC-12. The number in brackets behind each author describes the number of treated topics, and the filled circles in the plots represent the average value for the respective measure. Users 001, 014 and 025 are grayed out as they worked on few topics.

Figure 3.4 shows the user-wise distribution of the key figures from Table 3.3 in analogy to Figure 3.1. However, except for user 007 who generally produced long essays there are only few interesting findings to remark.

First, a correlation between the number of pastes and the number of used references is recognizable, i. e., the more pastes an author did, the more references were used in total. This might not sound surprising, yet one could have expected that dedicated users paste many short passages from only few key documents and include those passages carefully into their existing essay in favor of a high logical coherence.

Second, we expected users with comparatively few revisions to paste a large number of text passages from foreign documents. However, as can be seen in Figure 3.4, users 006 and 007 both needed only about 500 to 1,500 revisions to complete their essays. While these values are close together, the number of paste events differs greatly, and thus no correlation between number of revisions and paste events can be inferred. Speaking about user engagement, which is the matter of interest in Section 4.5, we cannot actually state whether a user was highly dedicated just by concentrating on the number of revisions or pastes.

3.4 Summary

This chapter provided an overview of the data basis that is used for the upcoming analyses. The Webis-Query-Log-12 was examined in detail and compared to its initial release. Moreover, some interesting characteristics could be revealed for both the Webis-Query-Log-12 and the Webis-TRC-12 that motivate the research questions to treat in Chapter 4.

4 Characterization of User Behavior

The following sections present several analyses that help to characterize and understand user behavior in exploratory search tasks. Each of the sections is organized as follows: First, the motivation for this particular analysis is presented along with our expectations. Second, the methodological procedure is described so that upcoming studies have a starting point from which they can begin their analyses. Third and last, the results will be presented and, whenever appropriate, be discussed with regard to earlier findings.

4.1 Visualizing User Interactions

Starting from scratch, the vast amount of available information in both logs is hard to grasp. A visual illustration of the logged interactions could help to better understand the temporal course of actions that users did during their exploratory search task. Further, these illustrations might reveal some (behavioral) patterns on a glance, which are worth to further examine.

Our visualization scheme should contain following information:

- Temporal order of interactions and their respective durations
- Physical sessions and their partitioning over different working days
- Development of essay length over time and its linkage to physical sessions

Turn over to Figure 4.1 for such a produced visualization of topic 29, which contains all of the information listed above.

Procedure

Since the Webis-Query-Log-12 itself does not contain information about physical sessions, these have to be determined first. For that purpose, the timespan between two consecutive interactions is calculated and a session break is inserted when this timespan exceeds some threshold. Experiments with different values show that a threshold of 15 minutes seems to be suitable as it produces sessions of reasonable length but on

the other hand not hundreds of sessions per topic. Note that before determining the physical sessions, a customization of the original log data is applied: Text-writing interactions are subdivided into blocks with a new cut-off time of five minutes,[1] i. e., if one text-writing interaction contains two consecutive revisions with more than five minutes between them, the text-writing interaction is split into two distinct text-writing interaction blocks.

Once the physical sessions are determined, each session can be visualized within one row in our visualization scheme. Since only the text-writing interactions have an exactly known end time stamp, this information has to be estimated for query and click interactions. For queries, we apply an arbitrary threshold of 60 seconds because we assume that a user would not stare on the result list for more than one minute without clicking any result. For clicked documents we determine the number of words that are contained in the document and compute the expected reading time with an assumed reading speed of 250 words per minute, which is an acceptable value for most people [DLDL65]. By adding this value to the known start time stamp of the click, the end time stamp can be computed. However, as mentioned before, these are only estimations and it is not ruled out that a user really spent that much time reading a short document or viewing the search result list that thoroughly. Therefore, the visualization should not show a hard cut in these cases as it is done for text-writing interactions but rather signify that it is unlikely that the user really worked all the time to the next logged interaction.

Lastly, two lines should be added to the illustration in order to show the development of essay length, one line displaying the exact and the other a simplified progression. For the latter, only sampling points at the end of each session are used, thereby providing an indication about how much the actions of a particular physical session contributed to the final essay.

Results

Figure 4.1 shows such an illustration for the interactions of topic 29, which was already depicted as an example in Section 3.3. Therefore, one can notice that the progression of the dashed green line (the exact essay length) perfectly matches that one from Figure 3.2 and just the orientation has flipped.

Each row stands for one physical session, and the horizontal dashed lines divide different working days from each other. In Figure 4.1, only the sixth working day contains more than one physical session. The number 52 to the left of the second row shows the

[1] During log creation, this threshold was set to 15 minutes in order to keep possibly errors from our assumptions about users' behavior as low as possible.

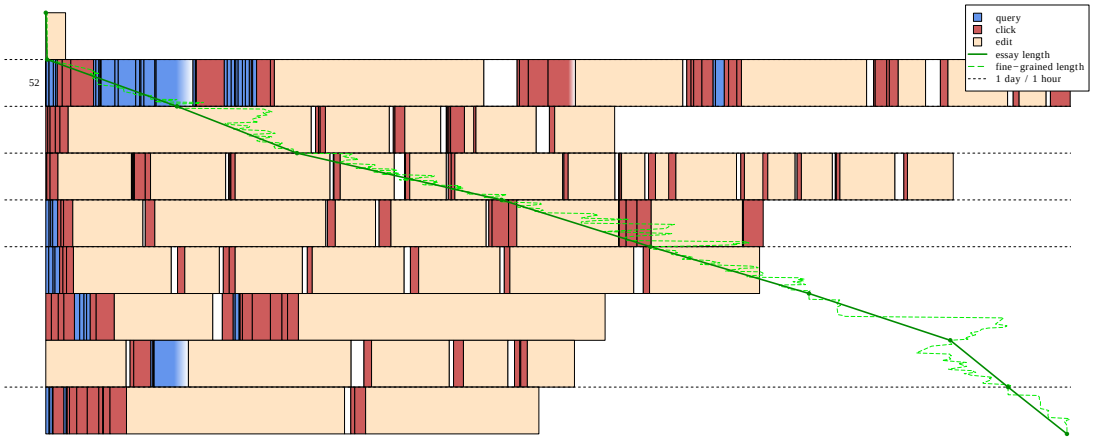


Figure 4.1: Visualization of all interactions that occurred during work on topic 29.

duration in minutes of the longest physical session. The beige blocks represent text-writing interactions, and the blue and red ones depict queries and clicks, respectively. Exact end time stamps are only known for text-writing interactions, so they are shown as a clear cut (cf. the white blocks after text-writing interactions in the second row). Color gradients in query and click blocks indicate that the user unlikely really spent that much time studying the result list or reading the document (see the description above). The two green lines show the progression of essay length over time, i. e., starting from the top-left corner where the essay consisted of no words at all, the essay grew up to its final length shown in the last row (last point is 100 %).

It can be clearly seen that text-writing interactions often directly follow on a click series. It seems that the user deliberately decided to learn and write about some particular aspect of the topic and visited a couple of documents in order to collect the needed information. Having finished one subtopic, the user changes over to the next. Compared to the number of clicks, the author submitted rather few queries and might be classified as a *clicker* (cf. Section 4.4).

Anyway, it is interesting to see that the author of topic 29 worked very purposeful, and that this continuous work extends evenly over the whole period of topic treatment. In contrast to this, Figure 4.2 shows another author whose working style seems to be not that organized. Starting with a couple of sessions in which the user 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. Potthast et al. call this *boil-down reuse*, whereas the pattern from Figure 4.1 is called *build-up reuse* [PHVS13a].

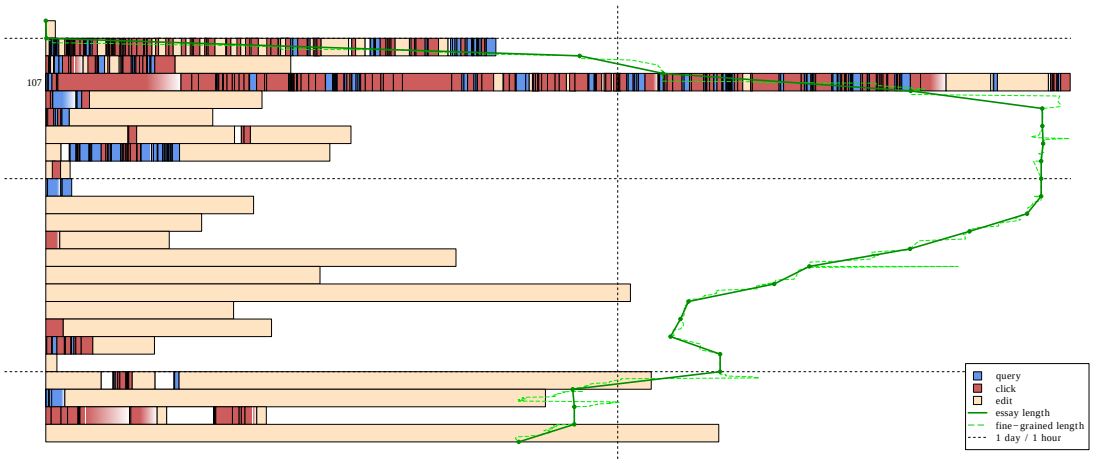
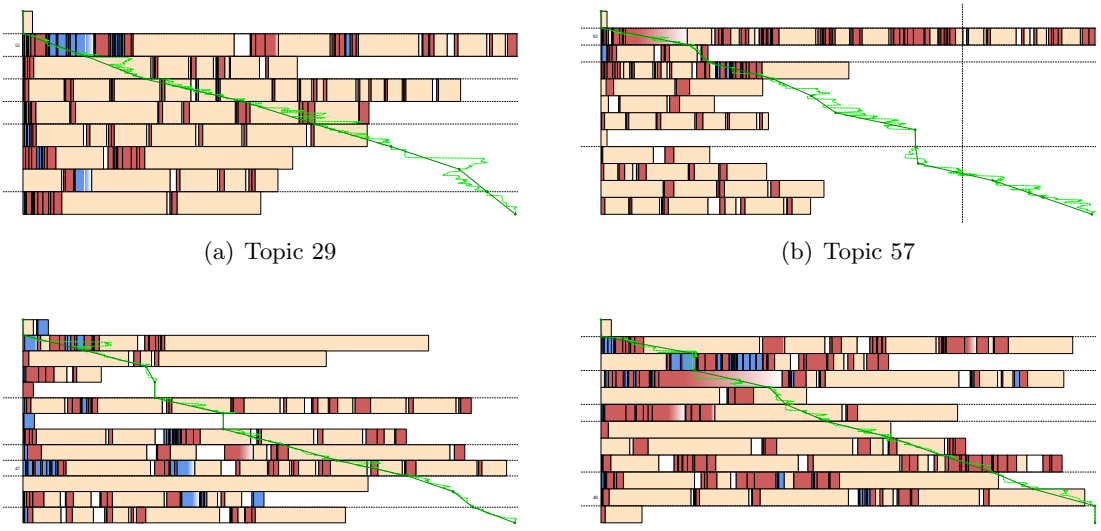


Figure 4.2: Visualization of all interactions that occurred during work on topic 27.

There is also another interesting detail about Figure 4.2: In the session before the last, a couple of clicks occurred, which are followed by very short text-writing interactions that influence the essay length only marginally. This can be observed quite often along many topics and different authors and was also recognized by Vakkari et al. [Vak10]. One explanation could be that users check their essay for possibly missing but important text passages from priorly selected sources.

Having inspected all of the 150 visualizations, it can be stated that each one provides a very fast and easy graspable overview of how a user treated some topic. User preferences like making use of references in a build-up or boil-down way can be identified on a glance and can even be used to draw inferences about the author just from looking at these plots. Figure 4.3, for example, shows four exemplary topics that were treated by user 005. The patterns described for topic 29 also hold for the other topics, i. e., the essay grows continuously and there are only few queries compared to the large number of clicks, which are often followed by longer lasting text-writing interactions.

However, as informative these visualizations are for a single topic, so diverse they are when examining all topics in an aggregation. Unlike our initial aspiration was, it is hardly possible to get an overview of all 150 topics at a time and to identify recurring patterns. Therefore, other mechanisms have to be found, which can draw more general conclusions from all of the log data and actually characterize users' behavior in exploratory search tasks instead of only visualizing it. Nevertheless, the generated visualizations for each of the topics can help clarifying obscurities and misunderstandings in further analyses, and we will return to them as needed.



(a) Topic 29 (b) Topic 57

(c) Topic 63 (d) Topic 114

Figure 4.3: Side-by-side comparison of interaction visualization for user 005.

4.2 Composition of Queries

It is typical for exploratory search tasks that users learn through reading documents and thereby extend or adapt their knowledge about the topic [EST+10]. Therefore, it is to expect that users' queries also develop over time, which should be proven by this section. Two analyses are conducted in this regard: First, it is shown which query terms occur when in the search process, and second, those terms are examined with regard to their presumed origin. We expect to clearly identify those terms that have been learned from previously visited documents.

Procedure

For the first analysis, we consider a visual check for all 150 topics again. For that purpose, stop words and any punctuation are removed from the lower-cased queries, which then get split up into their distinct terms. Highly similar terms, which have a cosine similarity of at least 0.85 and a Levenshtein distance of 1 at max, are conflated to one term. For each query in a topic, its composition of terms can now be visualized as dots in a plot showing the terms on the y-axis and the query number on the x-axis. See Figure 4.4 for a such a visualization.

Table 4.1: Origins of learned terms.

Task description	Clicked document	Result title	Result snippet	Initial knowledge
902 (24.3 %)	1,147 (30.8 %)	291 (7.8 %)	1,067 (28.7 %)	312 (8.4 %)

For the second analysis, we determine the possible origins from which a term can come from. These are, in the prioritized order:

- Task description
- A previously clicked document
- Document title of a previously shown but not clicked search result
- Generated snippet for a previously shown but not clicked search result
- User's initial knowledge

This order has been chosen because it describes the most promising time sequence of perceiving an unknown term: Before anything else, the user reads the task description and might pick up some task-specific vocabularies that will be of high importance for the whole search process. For each query in the interaction log, all terms are extracted and, in case the user entered them for the first time, assigned to one of the other possible origins, beginning with the interaction that precedes the currently examined query. As long as the term has not been found in a clicked document or in one of the shown results, the interaction list is traversed from the most recent interaction down to the very first one. This truly is a design decision as one could argue that it does not reflect the temporal order of term occurrence, i. e., the very first occurrence of a learned term might have been long time ago. Yet, we are interested in those documents and result lists that finally influenced the user's course and therefore decide for the most recent occurrence. If a term has not occurred during any of the prior interactions, it is classified as *initial knowledge*.

Table 4.1 shows the origins of all 3,719 distinct terms that appeared in the 150 topics. It can be seen that almost all terms are potentially learned during work on the topic, which is hard to believe. Indeed, most of these terms can be considered to stem from initial knowledge as a cross-check with the AoA (Age-of-Acquisition) set by Kuperman et al. shows [KSGB12]. This data set contains more than 30,000 terms and the average age (in years) at which U.S. citizens typically learn each of those terms. Only 388 out of the 3,719 terms have an average age of acquisition of 10 years or more. Choosing a threshold of 10 years again is a rather arbitrary design decision, yet words like *tribute* (10.4 years) or *fraud* (10.1 years) confirm that those terms are used comparatively seldom in every-day language. So, almost 90 % of the terms that are assumed to have been learned during the task are either from common parlance or not contained in

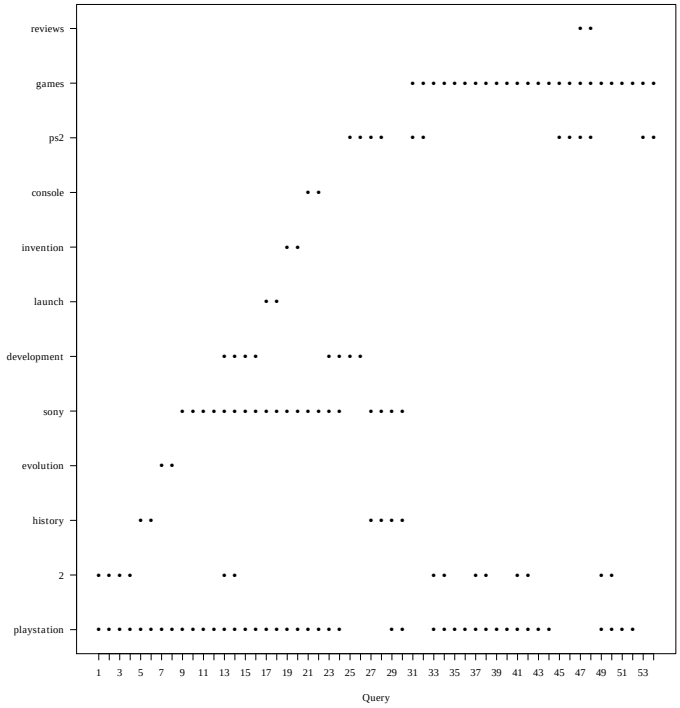


Figure 4.4: Query compositions for topic 29.

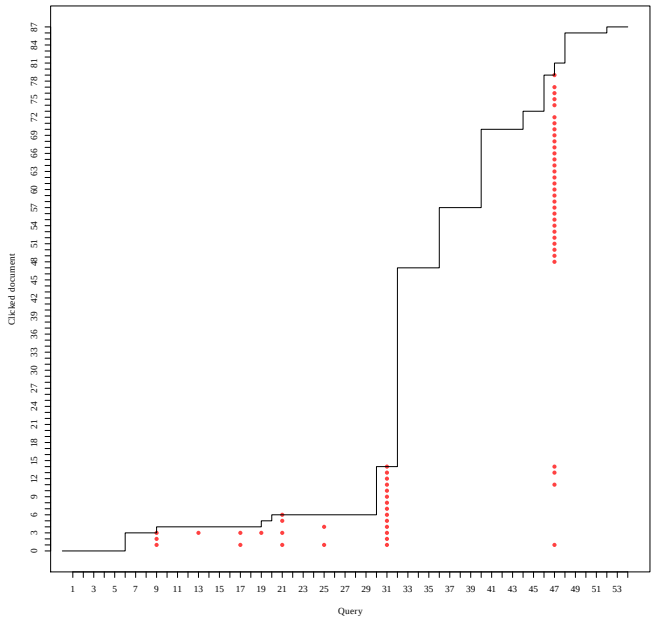


Figure 4.5: Learned terms and the documents of their origin for topic 29.

the AoA data set. Beyond this, terms that are classified as stemming from the user's initial knowledge are often misspelled words like *rpoviders* (instead of *providers*) or other artifacts like *siteenwikipediaorg* (remember that any punctuation is removed from queries for this analysis).

However, even though our classification seems to have several weaknesses, it clearly shows that the majority of used terms can be found in one of the priorly presented documents. Thus, we use the produced data to show in another visualization *when* in the search process a user introduced new terms and *where* they are likely to come from. Refer to Figure 4.5 for an exemplary result of this visualization.

Results

Figures 4.4 and 4.5 show the results of the two analyses for topic 29, which deals with the game console PlayStation 2. While the first figure is probably quite easy to understand, the second requires little more explanation. On the x-axis, one can see the current query number, which is the same as in Figure 4.4. The y-axis displays all clicked documents, and the staircase-shaped line depicts which click(s) happened as a result of which query. Here, in Figure 4.5 it is to see that the first three clicked results came from the result list of the sixth query, another click followed after submitting the ninth query, and so on. The dots indicate a new term in the query (x-axis) and all previously clicked documents that contain this particular term (y-axis). As Figures 4.4 and 4.5 are horizontally aligned, one can determine which exact term was introduced. If two or even more terms were introduced in only one query, each of the terms is represented by another color.

Of special interest in Figure 4.5 are the queries 31 and 47, which introduce the terms *games* and *reviews*, respectively. The occurrence of these terms is apparently not coincident as they are also contained in almost all of the clicked documents that were visited only a short while ago. However, the dots also indicate that these new terms occurred in documents that were visited a longer time ago, too; therefore, we cannot surely state whether or not the user really learned a new term only through the recent clicks.

However, such a vertical line of dots can be interpreted as a change of subtopic because the user ignored an often occurring term for quite a long time and then decided at some point to finally search for it. By means of topic 29, this change of subtopic is definitely true for the term *games*, as the prior terms *evolution*, *development* or *invention* clearly treat the history of the PlayStation 2.

Having examined all of the 150 topics, 70 of them contain clear indications for several subtopic changes, and another 18 topics even leave room for interpretation that some vocabularies must have been picked up on very recently visited pages.

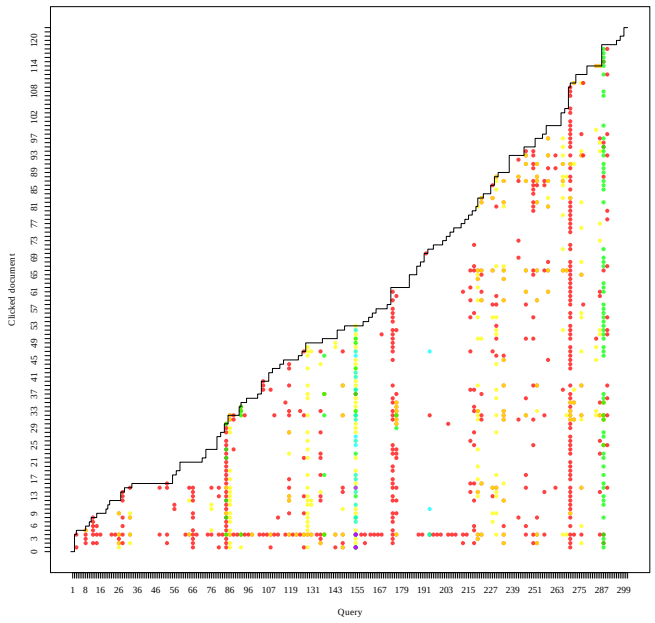


Figure 4.6: Learned terms and the documents of their origin for topic 133. Note that some of the points form a horizontal line, which is an outstanding property compared to almost all other topics.

Further, in two cases another notable pattern can be found. Unlike in Figure 4.5 and the visualizations of many other topics, the topics 42 and 133 do not only contain vertical lines of dots but also horizontal ones, as in Figure 4.6. Here, this pattern indicates that document number 4 must have been of special interest and the user got inspired in large parts by many of the vocabularies presented in the document, which is by the way a very detailed overview on the Declaration of Independence, the main theme of the treated topic.

However, coming back to the first part of this analysis, i. e., the term composition of queries, another oddity is directly visible from the visualizations: Many queries have numerous identical, immediate follow-up queries, i. e., no term has been added or removed. While this characteristic is obvious for those queries requesting more results than the initially presented top-10 documents, there are still 2,119 cases left in which the user requests either the same number of results or even fewer, i. e., 10 results instead of 100. For a typical example, consider the several groups of four dots in Figure 4.7, and notice that the query *chain link fence* is submitted ten times in a row (queries 67 to 76).

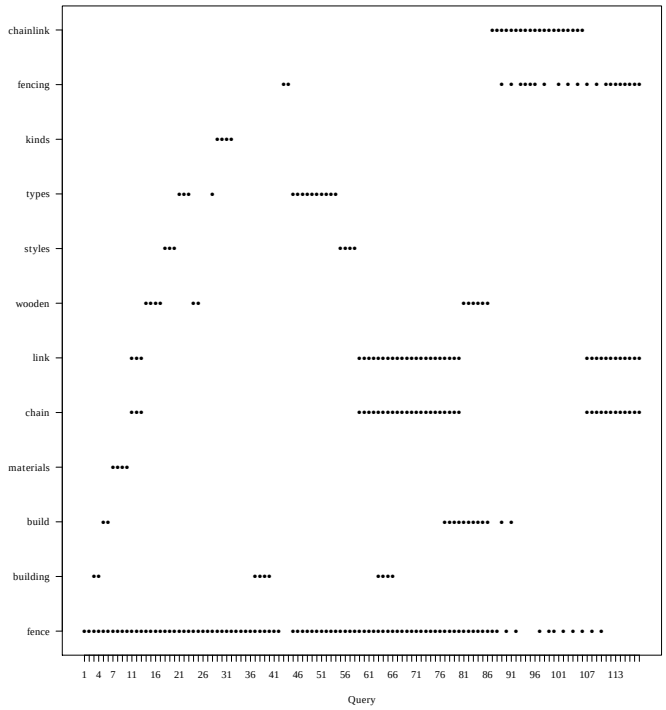


Figure 4.7: Query composition for topic 59.

Though not directly visible from the plot above but only from the data, the pattern here is always: 10 results, 100 results, 10 results, 100 results, and so on. So, the user returns four times to the same query after already having viewed 100 search results; the reason for this behavior is unclear.

We identify 480 out of those 2,119 cases to request the same number of results and a surprising quantity of 1,639 cases, in which the user returns to only 10 results although the preceding, identical query already showed 100 results. One possible explanation for this behavior could be that a physical session break interrupts the user's search activities. For example, a user submits an initial query, requests more results, perhaps clicks some of them, and then leaves the computer in favor of some other activity. When returning back to the task, the user starts with the same query that was submitted at the end of the last physical session. Yet, this is neither the case for the query *chain link fence* in Figure 4.7 nor for 1,898 other occurrences of this behavior, thus leaving 217 cases (10.2 %) in which indeed a session break occurred between the two queries. Another explanation concerns the retrieval quality of the used search engine, which is comparatively far below that of a commercial search engine. In 434 cases, fewer results were presented than originally

requested by the user (in 273 cases even no results at all were shown) and a user might have thought that re-submitting the query could produce more results. This turned out to be partially true, as in 132 out of the 434 cases actually more results were found when submitting the query a second time. Sometimes, also different results were shown for those duplicate queries that requested the same number of results again. Exactly the same results were shown in 198 of the 480 cases, whereas in 152 follow-up queries up to 20 % new results replace those from the priorly shown result list.

So, session breaks, unsatisfying results or perhaps even logging errors can explain at least some of the 2,119 cases, in which either the same number of results or even fewer has been requested. In Section 3.2 we discussed that only 3,538 distinct queries are contained in the log, which is little more than 25 % of all queries. A first analysis revealed that 6,727 out of 13,609 queries are identical follow-up queries requesting 100 instead of 10 results, so the existence of about 50 % of duplicate queries could be explained, yet it has to be clarified why the remaining 6,882 queries still contain so many duplicates (3,344 duplicates of 3,538 distinct queries). The found 2,119 cases from above can at least explain 63.4 % of them, and only 1,225 duplicate queries remain unexplained. We consider them to be *guiding queries*, which users return to from time to time for several reasons. First, the results of such a query can point to many directions for further investigations and a user might return to this query as soon as the work on one subtopic is finished. Second, guiding queries can serve to keep track of the main theme at any time and keep the user on course. And third, users might bring recently acquired knowledge into line with older knowledge structures and therefore want to return to previously seen documents. Some very typical guiding queries can be seen in Listing 4.1. They all have in common that they basically just reflect the main theme of the task.

```
discovery channel
computer keyboards
culpeper cemetery
jax chemical company
yellowstone national park
```

Listing 4.1: Typical guiding queries.

To sum up, in this section we showed why the Webis-Query-Log-12 contains so many duplicate queries, we evaluated the reasons for their existence and concluded with an explanation for those queries that users often return to from time to time. Furthermore, our analyses demonstrated that almost all used terms from users' queries could have been potentially learned during topic treatment, but a more sophisticated differentiation between terms from initial knowledge and actually learned terms has to be applied in order identify *key documents* that influenced the user's course.

4.3 Used References

Another way to identify such key documents is to analyze those documents from which users reused text passages in their own essays. If it was possible to find patterns in the usage of references, one could actively support users in their search task, for example, by query expansion [Eft96] based on those key documents. Or, alternatively, one could rank documents higher that were considered being useful by other users.

Procedure

We are especially interested in the differences between 'normal' clicks and reference clicks. Therefore, basic statistical analyses already should reveal if there are trends to see in the data or not. Possibly discriminating features are:

- Time spent for reading a document: We expect this to be longer on reference documents than on non-reference documents.
- Document length: Documents containing more content are naturally likely to contain more useful text passages that a user can plagiarize from.
- Duration of text-writing interaction after visiting a reference document: This should be a little longer for reference clicks as users might paste and rewrite content from that document or at least adapt their essay in course of their recently acquired knowledge.

Results

Table 4.2 summarizes the statistics. It shows that all of the aforementioned expectations are met; however, the differences are only of marginal size. Note that those differences are highly significant as shown by a Mann-Whitney U test (data is not normally distributed, so a Student's t-test is not applicable).

The values from Table 4.2 certainly elicit more questions. Recall Section 3.2, for example, which showed that 661 clicked documents have been viewed for at least 150 seconds, which is six times longer than the median reading time for reference documents. Therefore, it might be interesting to examine whether these documents have an especially high value for users. Surprisingly, it turns out that only about one third (240 documents) are reference documents and the other two thirds (421 documents) are not. So, reading time on its own is not a good indicator for identifying key documents. Even the ratio of expected reading time to actual reading time cannot help in this regard. Our data contains 5,039 clicks on reference documents (since only 2,761 references were used overall,

Table 4.2: Descriptive statistics about differences between reference and non-reference documents.

Analysis / Group	Q1	Median	Average	Q3	Significance
Reading time (seconds)					
Reference	8.0	25.0	44.0	52.0	$U = 26{,}206{,}838.0,$
Non-reference	7.0	20.0	38.3	43.0	$z = -8.4,\ p < 0.001$
Document length					
Reference	438.0	645.0	981.1	1,010.0	$U = 9{,}841{,}088.5,$
Non-reference	376.0	544.0	925.2	897.0	$z = -10.2,\ p < 0.001$
Duration of following text-writing interaction (seconds)					
Reference	40.0	135.0	357.4	418.0	$U = 1{,}366{,}246.5,$
Non-reference	15.0	64.5	279.8	305.0	$z = -10.1,\ p < 0.001$

each reference document was clicked 1.8 times on average). Out of these 5,039 clicks, users spent in only 277 cases more time reading the document than they are expected to. Again, the expected reading time is calculated by the number of words divided by an assumed reading speed of 250 words per minute. However, we are still dealing with the issue that no exact reading times have been recorded. We *assume* a reading interaction to have been finished as soon as the next interaction starts, but it is definitively possible that users kept a document open in a background tab and spent much more time reading it than we can see from the log data. Knowing the exact reading times could definitely help in our analyses and might lead to a well-discriminating feature for identification of key documents. So, one strong recommendation for possibly upcoming log crawlers is to implement mechanisms that allow more accurate tracking of actual user interactions across different tabs and browser windows.

It is also interesting to see, which interactions follow such a reference click. As one might have expected, users most likely continue with a text-writing interaction (2,613 of 5,039), closely followed by a new click (1,953 of 5,039), and only in comparatively few cases users submit a new query (473 of 5,039). Especially in respect of the origin of such reference clicks, the number of following click interactions is notable as 786 reference clicks stem from click trails that already contained one or more reference clicks. This indicates that users are likely to find more useful reference documents as soon as they found one. So, if it was possible to accurately identify key documents, a recommendation system could be built that follows all links from a document and ranks those linked documents with regard to the user's last query.

Despite the weaknesses described above, we try to identify such key documents in a last step of this section. The approach is rather low-leveled as are the data we can use for it. For each click in a topic, the document's length is determined, the time spent for reading it, as well as the time spent on the next text-writing interaction. These three features are then compared to the median values for reference clicks from Table 4.2. If any of

Table 4.3: Confusion matrix for the key document identification algorithm.

		Log data	
		Reference	Non-reference
Algorithm decision	Reference	TP = 2,386	FP = 3,867
	Non-reference	FN = 2,766	TN = 7,679

those features reaches some specific percentage of the corresponding median value, the condition for this feature is set to true. If two out of three conditions are fulfilled, the clicked document is assumed to be a key document. As an example, consider a click on a document containing 580 words, which was viewed for only 4 seconds before going over to a text-writing interaction that lasted 180 seconds. For an assumed threshold of 80 % two out of three conditions are fulfilled (document length: $580 \div 645 = 89.9\,\% > 80\,\%$, reading time: $4 \div 8 = 50\,\% < 80\,\%$, time for text-writing: $180 \div 135 = 133.3\,\% > 80\,\%$) and the clicked document is assigned to be a key document.

In order to determine an appropriate threshold value, we simply test all percentages from 80 to 200 and decide for that percentage with the highest $F_{0.5}$-score. The $F_{0.5}$-score is chosen because we try to optimize towards precision rather than recall. That means that the number of documents erroneously classified as key document (false-positive) should be kept as low as possible, thereby potentially rejecting many actual key documents. The best result can be achieved for a threshold of 85 %, reaching an $F_{0.5}$-score of 0.396. Table 4.3 presents the confusion matrix for the decisions of our simple key document identification algorithm. It shows that only 2,386 out of the 5,039 (TP+FN) key documents were correctly identified. On the other hand, 3,867 documents are erroneously identified as key document, which is even more than the correctly identified. To conclude, an algorithm for key document identification based on those weak values that we used does not perform very well, and better metrics should be found.

4.4 Searching Strategies: Clickers and Finders

This section concentrates on elementary differences in users' searching strategies. As already observed in Section 3.2, some users submit only few queries but follow long click trails, and others submit a variety of queries but seldom click one or even more of the presented results. We call users following either of these two searching strategies *clickers* and *finders*.

Table 4.4: Queries and clicks by user 006 between consecutive reference document clicks.

	Min	Q1	Median	Average	Q3	Max	Significance
Queries	0.0	0.0	4.0	4.5	6.0	22.0	$U = 2{,}504,$
Clicks	1.0	1.0	1.0	1.8	2.0	6.0	$z = -3.317,\ p < 0.01$

Procedure

To distinguish between clickers and finders, we simply count the number of queries and clicks that are performed until a reference document is clicked. It is not important how many queries and clicks occurred overall but only how many of them occur between two clicks on reference documents. For the seven topics treated by user 006, for example, we collect the statistics shown in Table 4.4. Since the number of queries significantly exceeds the number of clicks user 006 can be assigned to the group of finders. Considering the median values, user 006 submits four queries in order to find one reference document but clicks only one document: the reference document itself.

Results

The analyses for all users indeed reveal that the two priorly mentioned groups of clickers and finders can be found. Users 005, 007, 020, 021 and 024 can be identified as clickers, and users 002, 006, 017 and 018 are identified as finders. Users 001, 014 and 025 again are not evaluated for they worked on too few topics (2, 1 and 1 topics, respectively), yet the trend shows that they tend to be clickers.

In order to confirm the differences between clickers and finders, we compute some basic key figures that are listed in Table 4.5 and illustrated in Figure 4.8. Except for the number of clicks, which is fairly high for finders, all differences between both groups are highly significant as shown by a Mann-Whitney U test (again, the data is not normally distributed and consequently a Student's-t test cannot be applied). The fairly high number of clicks in the finder group simply seems to depend on the number of queries submitted (cf. the similar shapes of boxes and whisker extents in the first two plots

Table 4.5: Oppositions of median values for the two groups clickers and finders.

	Clickers	Finders	Significance of difference
Queries	47.0	107.0	$U = 1{,}058.5,\ z = -6.148,\ p < 0.01$
Clicks	102.5	79.5	$U = 2{,}074.0,\ z = -2.136,\ p < 0.05$
Pastes	39.0	19.0	$U = 1{,}361.5,\ z = -4.952,\ p < 0.01$
References	17.5	15.0	$U = 1{,}876.0,\ z = -2.921,\ p < 0.01$

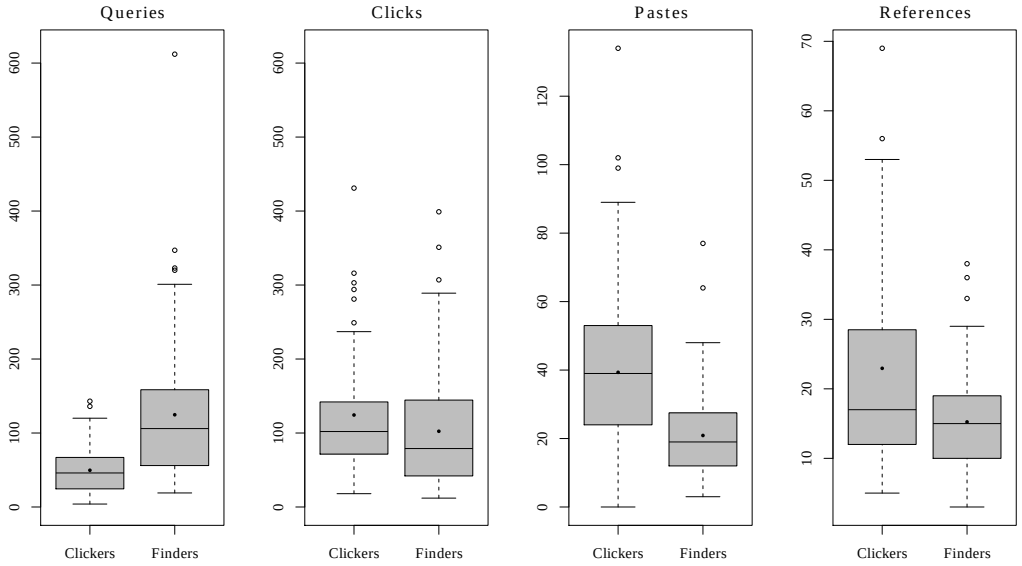


Figure 4.8: Oppositions of some key figures for the two groups clickers and finders.

in Figure 4.8; the ranges of y-axes have been aligned for this reason). After all, the distributions of clicks for both groups differ not as much as the distributions of queries, pastes and references. This leads to the assumption that users in exploratory search tasks consume –independent of the applied searching strategy– some particular quantity of informative content before considering themselves to have learned enough. So, it is not particularly good or bad to belong to the group of either clickers or finders.

Further, in Figure 4.8, it is notable that clickers paste about twice as many times as finders 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 finders seem to select their references little more carefully than clickers do. The reason for this behavior remains unclear and is left for future work.

4.5 User Engagement

Now that it is clarified that different searching strategies are neither to judge as good nor as bad, are there any other indicators for *user engagement* to find in the log data? User engagement here means how much effort a user puts into treatment of the task,

which can be a valuable information for a search engine. For example, a truly dedicated user might be interested in additional resources that go far beyond the original query intent, whereas a user who works only unwillingly on a task might be only interested in overview pages without having to deal with too many details.

So, how can 'lazy' users be characterized? Or at least, what characterizes a poorly treated topic? Do only few queries or clicks occur? Are there only few edits, used references, etc.? It is not this section's claim to develop and present a universally valid measurement for user engagement, but as we are equipped with a comprehensive base of 150 treated exploratory search tasks, at least a ranking within those topics should be produced.

Procedure

In this regard, the following nine features are extracted from each topic:

1. Number of distinct queries
2. Number of distinct clicks
3. Number of pastes
4. Number of used references
5. Total working hours
6. Number of physical sessions
7. Number of handled subtopics[2]
8. Total time spent for reading documents
9. Total time spent for text-writing interactions

All of these features have an impact on the overall invested time, which (for now) should be our indicator for user engagement. Note that not only the fifth feature (total working hours) is taken for this measure but all of the presented ones. Therefore, even a task that was completed in only two hours can be rated as sufficiently treated if it has, for example, many references or a large number of handled subtopics.

In the next step, a ranking along all topics is produced for each feature individually, and each topic gets a score depending on its rank. For example, topic 133 contains the most distinct queries along all topics and thus obtains 121 points (it is not 150 because 29 topics share the same number of distinct queries and therefore obtain the same score). For

[2] To determine the number of handled subtopics, an algorithm for search session and mission detection by Hagen et al. [HGBS13] is applied. The number of different mission IDs corresponds to the number of handled subtopics.

the feature *distinct clicks*, topic 133 is only on rank 18 and obtains 77 points. This is done for all features, and in a last step, all feature scores of a topic are summed up. Sorting the topics by their overall score in a descending order produces the final ranking.

Results

An excerpt of this ranking is shown in Table 4.6, which presents those 10 topics that have been treated with the most and the least engagement, respectively. It is remarkable that 9 of the top-10 topics were treated by user 002, who seems to have worked really dedicated in almost all tasks, whereas users 006 and 024 seem to have worked with little enthusiasm on their topics.

To identify the most and the least dedicated users, we simply compute the average for each user in order to bypass the different numbers of treated topics. It turns out that user 002 indeed belongs to the most dedicated users with an average score of 403.5 but is slightly outperformed by user 021 with an average of 404.8. Note that user 002 worked on 33 different topics and the range of scores is widely distributed, whereas user 021 worked on only 12 topics, which all achieved quite high engagement scores. The least dedicated users in our collection are user 006 and user 020 with an average score of 141.9 and 188.6, respectively.

In order to see which of the features are especially discriminating, we compare the median values for all features between the two most and least dedicated users. Figure 4.9 shows the nine features and how far the more and less dedicated users lie above or below the median value, which is computed over all 150 topics. It can be seen that the number of distinct queries actually is not a good indicator for user engagement as the less dedicated users 006 and 020 both slightly lie above the median, whereas the most dedicated user 021 is even located below this value. As we can see in the figure, the number of distinct

Table 4.6: Topics ranked by the engagement they were treated with.

Rank	Topic	User	Score
1	Topic 58	User 002	551
2	Topic 53	User 002	538
3	Topic 110	User 002	524
4	Topic 13	User 021	523
5	Topic 67	User 002	503
6	Topic 27	User 002	499
7	Topic 49	User 002	498
8	Topic 144	User 002	493
9	Topic 10	User 002	484
10	Topic 22	User 002	479
...			
141	Topic 89	User 007	115
142	Topic 32	User 002	113
143	Topic 3	User 024	111
144	Topic 124	User 018	104
145	Topic 23	User 024	89
146	Topic 147	User 006	74
147	Topic 116	User 006	63
148	Topic 43	User 020	62
149	Topic 146	User 006	45
150	Topic 21	User 024	40

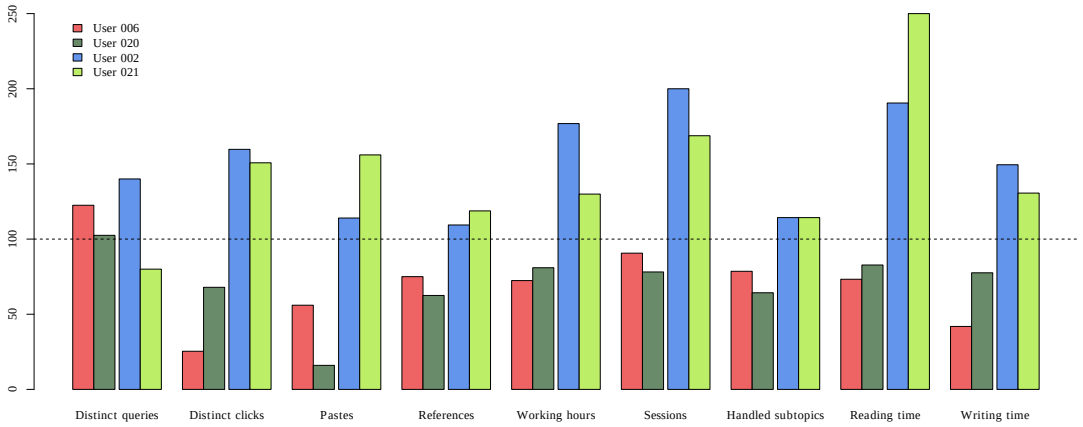


Figure 4.9: Comparison of median values for the nine features. Users 006 and 020 are the least dedicated, users 002 and 021 the most dedicated users in our data set. The horizontal line at 100 % shows the median value over all 150 topics.

clicks (and consequently the reading time) as well as the number of pastes seem to be especially valuable indicators for user engagement. The large differences in reading time might imply that the tasks were more difficult than others and therefore this feature is not actually convenient for measuring user engagement. Yet, since users were allowed to choose the tasks they like to work on, the choice of a comparatively difficult task may be a further indicator for highly dedicated users.

Figure 4.9 demonstrates that despite of the simplicity of our user engagement heuristic, clear trends can be found between more and less dedicated users. However, retrospectively to Section 3.3, we cannot actually state that the number of used references on its own allows direct inferences on user engagement.

Moreover, it is important to note that the engagement ranking presented above does not imply any conclusions about the quality of the essay itself but only about the effort that users spent for writing it. The quality of the essay has to be determined in a separate step, for which we do not come up with a solution in this thesis. However, having such a measure at hand would allow to easily prove whether our user engagement heuristic correlates with the task outcome (cf. [VH12]) or not.

4.6 Types of Authors and their Writing Styles

As a side note from Section 4.5 we also recognize differences between both groups in their text writing behaviors, namely the time spent for writing and the number of pastes. It seems there are different writing styles along the authors, which will be evaluated in this section.

Indeed, Potthast et al. found different patterns in essay creation and distinguished between *build-up* and *boil-down* strategies [PHVS13a]. The first describes a rather continuous essay growth, i. e., mostly short text passages are continuously added and almost immediately adapted to fit into the essay structure. Recall Figure 3.2 and note how the essay length line steadily grows. On the other hand, boil-down authors first gather as many possibly useful text passages and paste them one after another into their essay. Afterwards they begin to reorganize those passages, thereby dropping a considerably large quantity of text fragments resulting in a shrinking essay. Figure 3.3 shows a typical representative for this type of writing style. Here, even two information gathering phases can be clearly seen: one at the very beginning of topic treatment comprising 14 paste events, and a second one at about the middle of the text-writing process when the essay grows again though only three paste events happen (or four, depends on interpretation).

Potthast et al. manually identified 65 build-up essays, 65 boil-down essays, 19 others in which both styles are mixed, and 1 essay which produced some error during evaluation [PHVS13a].[3] A cross-check of their decisions reveals that in a couple of cases none of the patterns can be applied unmistakably, even not the mixed-pattern. For example, consider the blue line, which depicts the essay length without HTML residues, in Figure 4.10. Here, almost no evidence for any trend of development of essay length is visible at all. Yet, it seems obvious that the user had several information collection phases in which text passages from diverse source documents have been pasted into the essay, which then are reorganized in the follow-up phase. Since this behavior is typical for the boil-down pattern, the essay from topic 48 has a slight tendency into this direction, which was undetected by Potthast et al. as they have not taken the regularity of paste events into consideration.

However, cases like these and the issue that the classification was based only on visual inspection instead of a precise measure motivate us to revisit the classification of different writing styles. As demonstrated with the aid of the essay from topic 48, we argue that the general tendency about essay growth truly is one important indicator for the applied writing style but that we also need to consider the regularity of paste events as another important indicator for a valuable classification.

[3] The specified values stem from the authors' original data but cannot be found in the publication where only the combined values for Batch 1 *and* 2 (cf. the two different scenarios described in Section 3.1) are presented.

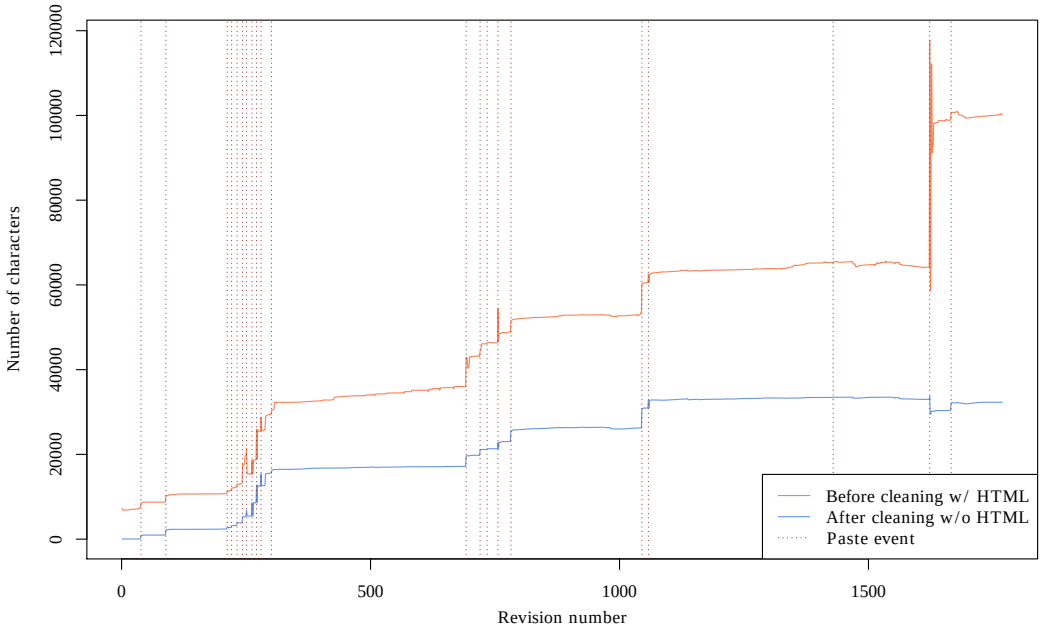


Figure 4.10: Essay completion graph with paste events for topic 48.

Procedure

In order to determine a measure for the regularity of paste events, we decide for the following procedure: First, the paste events are obtained by the method described in Section 3.3. Second, the number of revisions that lie between each consecutive pair of paste events are collected. For example, if an essay was completed within 50 revisions and there occurred paste events in the revisions 10, 22 and 40, we would have collected the list of values $\langle 10, 12, 18, 10 \rangle$. Note that the number of revisions between start of work and first paste event appears in this list as well as the number of revisions that lie between the last paste event and essay completion. This is important because we want to compute the paste regularity over the full working time and not only parts of it, i. e., an author who pastes very regular in each 10th revision at the beginning of the working time but then does not paste any more content in the following 3,000 revisions up to essay completion actually does not paste regularly. In the third step, the variance of the determined list is computed. A low variance means that the paste events are rather equally distributed over the essay revisions, whereas a high variance indicates that a user pasted very irregular. While this is a very simple approach, the results are of comparable quality to other approaches like calculating the root-mean-square deviation of percentages for essay completion and paste events.

To obtain a measure for development of essay length we simply check for all subsequent revisions whether at least one full word was added or removed. If either is the case, a counter is increased. It does not matter *how many* words have been added or removed; only the trend matters. Revisions that contain a paste event are ignored. At the end this might result in 400 revisions in which content was removed and 600 in which content was added. So, it may be stated that the essay tends to grow, as 60 % of the relevant changes lead to a longer essay. Yet, each of the essays has to grow naturally as 5,000 words have to emerge somehow. Therefore, a value of 60 % actually is an indication for a boil-down pattern (cf. the x-axis in Figure 4.11).

Results

For each author and topic both values, i. e., the trend of essay length development and the paste regularity, are calculated and then arranged in a 2D plot. Figure 4.11 shows the resulting plot. Each of the symbols depicts one essay, showing the trend of essay growth on the x-axis and the paste regularity on the y-axis. Different symbols (and colors) indicate different users, thus revealing trends for each author's writing style.

The x-axis ranges from about 50 % to almost 100 %, meaning that those essays on the right-hand side of the plot hardly ever have been revised but that content was added permanently. Interestingly, the two users 006 and 020 who are isolated from all other authors by reaching an essay growth of at least 85 % in all of their essays, were considered to be the least dedicated users in Section 4.5. This may indicate that those essays only form a lethargic sequence of various plagiarized text passages, probably without too strong coherence under another. Essay growths that range from 65 % to 85 % can be considered to be the rule rather than the exception. Those essays have a well-balanced proportion between text insertions and removals. Essays with a growth below 65 %, here especially those by user 002, are those that were considered to show up the boil-down pattern in the classification scheme by Potthast et al. [PHVS13a]. For user 007 one great advantage of our new taxonomy over the old classification procedure becomes visible: Potthast et al. identified the writing styles through visual inspection, which sometimes can be misleading when viewing a plot in its entirety but not in its details. For example, topic 7 was classified as build-up essay, which is absolutely comprehensible when viewing Figure 4.12 on a first glance. However, the truth is that the essay chiefly grew through pastes and actually shrank between the paste events. Figure 4.11 correctly uncovers this property and the essay of topic 7 shows up at 50 % on the x-axis, thus rather indicating a boil-down strategy.

Yet, as can be seen on the y-axis, a boil-down pattern does not necessarily mean that a user has a sloppy working style. In fact, Figure 4.11 shows for user 007 that the paste events occurred very regular as all symbols appear at the upper edge of the plot. In this

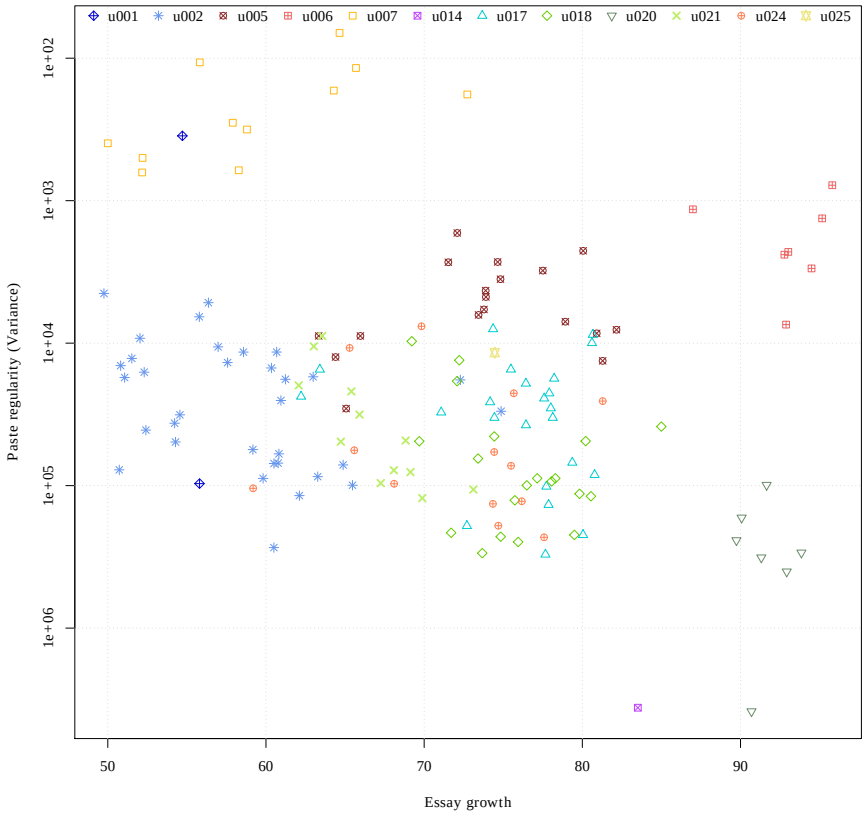


Figure 4.11: Writing styles for each author and topic.

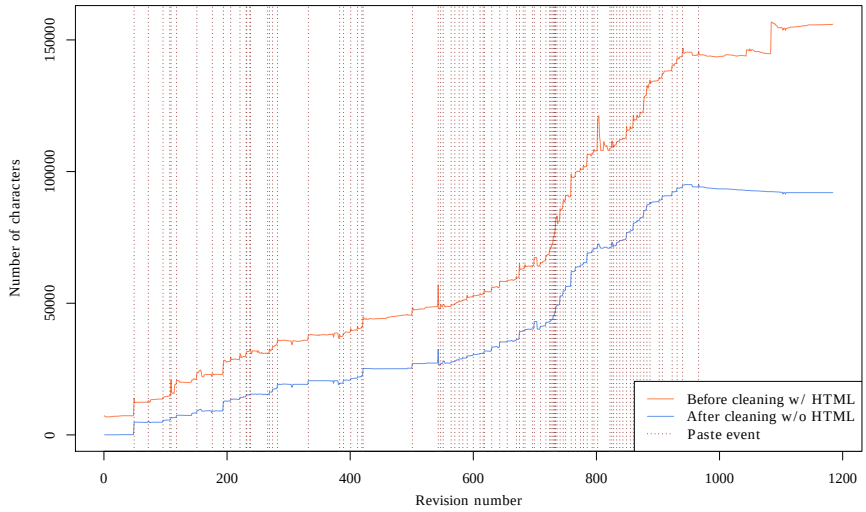


Figure 4.12: Essay completion graph with paste events for topic 7.

regard, note that the y-axis is flipped, i. e., the smallest variance is shown at the top of the plot and the highest variance at the bottom. This is done to map our natural understanding of a *high paste regularity*, which requires a low variance, to the *upper part* of the plot. Note also that the y-axis is displayed on a logarithmic scale as differences in variance might become quite large values.

The probably most remarkable outcome of Figure 4.11 is that different authors can be visually clustered by their writing strategies. In many cases, the data points for all essays of an author are located in one specific area of the plot. For example, user 002 has a moderate paste regularity and could be classified as boil-down writer, whereas user 006 follows a strict build-up pattern and has a rather high paste regularity compared to most other authors. It is left for future work to develop an algorithm that can automatically assign a given essay to one of the users.

Figure 4.11 proves not only the existence of different writing –or better– working styles but also that people evolve characteristic traits towards their own working style, which they hardly can cast off. Further evidence for characteristic traits of each user can be found in Appendix A.

4.7 Comparison of Working Phases

As a last analyses in this thesis, it should be examined whether different working phases in exploratory search tasks can be identified. Do users submit more queries in the early phases? Are the reading phases located in the middle? And do text-writing interactions form the major part in late phases? If any patterns could be found, this may facilitate a search engine to support users in their respective working phases. In the early phase, for example, a search engine could present not only search results for the submitted query but also suggest shortcut queries [BCC+09] that helped other users finding relevant documents on the treated topic. While this could be helpful in the early phase to quickly acquire an overview of different aspects of a topic, it might not be desirable anymore in a later phase in which a user is only interested in a particular detail.

Procedure

For the reason of simplicity, we subdivide each topic into only three and not more working phases, just as already suggested above: early, middle and late. Subdivision is not done by splitting up the interaction list into three bunches of equal size but into bunches of equal natural time. This means that the *actual working time* (the sum of each session's duration) is divided by 3 and all interactions that fall within a certain third are collected. Note that this can result in interaction lists of different sizes. Then,

the percentage of the four features *queries*, *clicks*, *text-writing interactions* and *pastes* is calculated for each of the working phases. For example, if 30 queries out of 60 appeared in the first third of natural working time it is 50 % for the early phase, and so on. Values for one feature sum up to 100 %.

Results

For each user, one plot is generated containing all topics that were treated by this user. Figure 4.13 shows the percentages of the four features over the different working phases for user 017. Each colored line represents one treated topic, and the same color is used in all four plots for the same topic. The general trend shows that the number of queries,

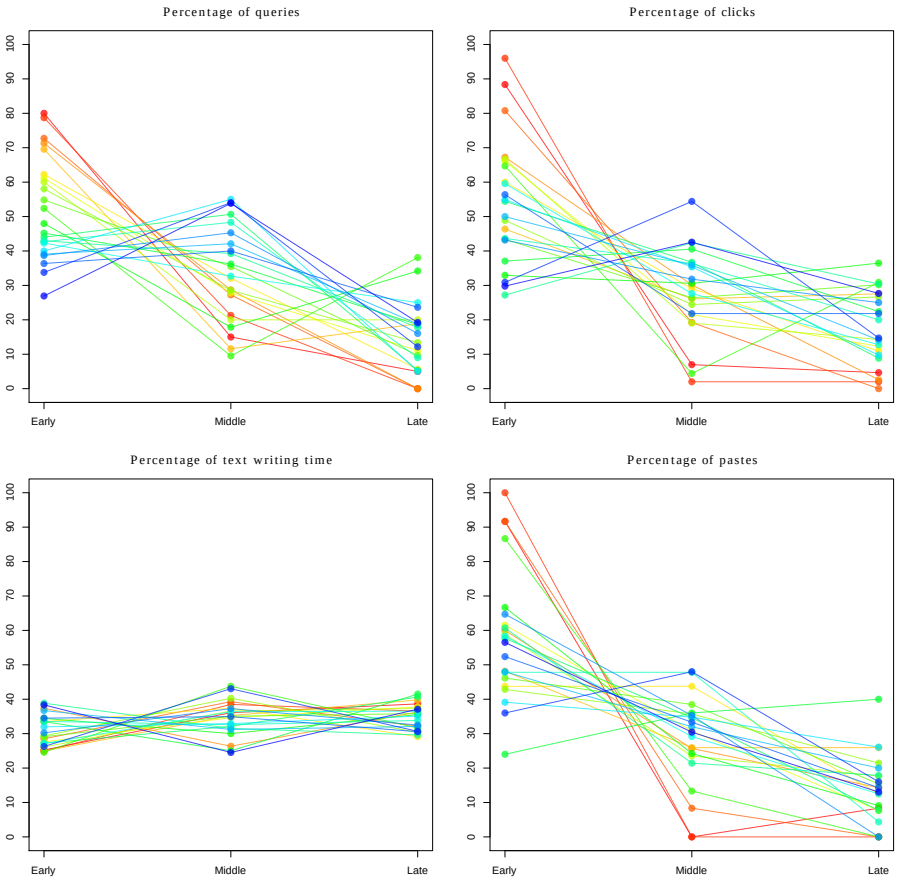


Figure 4.13: Comparison of working phases for user 017.

clicks and pastes tends to fall, while the number of text-writing interactions increases in the later working phases. Yet, the different pathways of the lines do not actually reveal any particular pattern, especially in the case of the top-left plot, which shows the percentage of queries. In two cases, the lines have a V shape indicating that for some reason user 017 submitted a couple of queries right before essay completion. Recall that we already discovered this behavior in Section 4.1 and provided the possible explanation that the user might have checked the essay for possibly missing text passages from previously clicked documents. Another conceivable explanation is that this is a fact-checking mechanism, i. e., user 017 verified the facts written in the essay right after its completion. In contrast, the bluish and turquoise lines rather have an A shape, which suggests that user 017 perhaps submitted several tentative queries in the early phase in order to explore the structure of the information space [WR09] and in the mean phase submitted more sophisticated queries asking for more detailed information.

It is also remarkable how large the distribution of values for one working phase can be. Consider the top-right plot, which shows the percentage of clicks over the working phases. In some topics, about 30 % of all clicks happened in the early phase, which seems reasonable, and in some other topics even more than 80 % of the clicks have already happened when the first third of time has just passed. It is even more remarkable that the percentages of clicks and pastes approximately correlates, which can also be found along most of the other authors. This might be an indication for a constant factor between clicks and pastes, e. g., on average each 10th document is a useful one. It does not seem that users in exploratory search are able to improve their precision (in terms of queries and clicks needed to find a desired information) just in course of the time spent on the topic. However, since an ascertained statement can hardly be made just from viewing the plots, it is left for future work to show whether or not users can improve their efficiency in the later phase of topic treatment.

Furthermore, we are interested in a comparison of potential differences in organization of the different users' working phases. Therefore, the median values for each author and working phase are determined for each feature. These values are then combined into one plot, which might reveal patterns that are typical for specific users. Yet, as Figure 4.14 shows, the general trends that were detected for user 017, i. e., queries, clicks and pastes decrease over time and text-writing interactions increase, also hold for almost all other users. This makes it hard to differentiate between users just on the basis of these lines.

4.8 Summary

In this chapter we conducted various analyses, ranging from rather low-level approaches like the composition of queries to more complex ones like a first draft of an algorithm

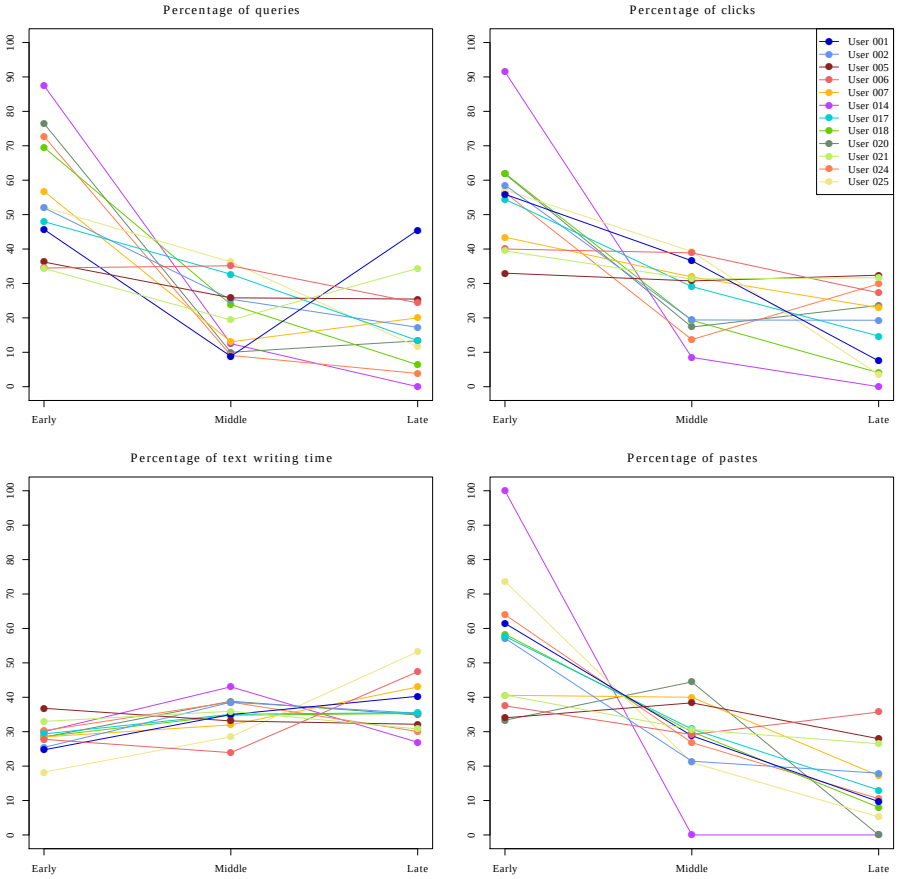


Figure 4.14: Comparison of working phases for all users.

that determines whether a clicked document is potentially useful or not. It was also confirmed that the build-up and boil-down text reuse patterns established by Potthast et al. [PHVS13a] can be found along the essays. However, we argue that a more sophisticated taxonomy is needed, which was proposed in Section 4.6. Interestingly, it turned out that data points for one user are often located in some particular area of the plot. This and other major outcomes of this work are summarized and discussed in Chapter 5.

5 Summary and Discussion

This work examined user behavior in exploratory search tasks. A summary of the most important key findings is provided in Section 5.1, an outlook on upcoming work is presented in Section 5.2, and Section 5.3 will conclude this thesis with a short résumé.

5.1 Main Contributions

In order to characterize user behavior in exploratory search tasks, we conducted analyses on two logs, namely the Webis-Query-Log-12 and the Webis-TRC-12. Since the Webis-Query-Log-12 was revised in large parts, Section 3.2 carefully described and examined this revised interaction log. The major contributions compared to the initial release are (1) the newly introduced XML format and (2) the clearance of various errors including missing log data, which was recovered in the course of this thesis. Further, general trends were detected by applying descriptive statistics which motivated and lead to the majority of analyses conducted in Chapter 4.

In Section 4.1, a visualization scheme has been proposed that provides a fast and easily graspable overview of all interactions that occurred during the treatment of a topic. Although this visualization is quite informative for each single topic, it is hard to draw general inferences about user behavior in exploratory tasks.

The subsequent Section 4.2 analyzed the term composition of users' queries and identified 18 topics for which it is likely that the respective authors have actually learned keywords from recently visited documents. At least, it has been proven that almost all terms could potentially stem from previously clicked documents and that a more sophisticated technique is needed to distinguish between terms from initial knowledge and those that actually have been learned during topic treatment. Further, Section 4.2 evaluated the reasons for often occurring duplicate queries and exposes the existence of *guiding queries* that users return to from to time in order to (1) pick up again an earlier search result list, (2) keep track of the main theme, or (3) bring recently acquired knowledge into line with earlier knowledge structures.

Section 4.3 examined apparently useful documents, i. e., documents from which content was reused in the authors' essays. Based on three features (reading time, document

length and duration of next text-writing interaction), a basic algorithm has been developed that should determine whether a clicked document is likely to be a useful one or not. In the evaluation, it turned out that this algorithm performed rather poor with an $F_{0.5}$-score of 0.396. Since we expect to achieve significantly better results if the exact reading times were known, we recommend to implement mechanisms for future log crawlers that can keep track of users' multi-tabbing behavior.

In Section 4.4, the existence of two different searching strategies has been proven and we divided the authors into two distinct groups, which we called *clickers* and *finders*. Finders submit significantly more queries than clickers and often click only one or two results of the result list, if any at all, whereas clickers tend to click more results and follow rather long click trail paths. Yet, it has been shown that neither of the applied strategies is by default good or bad as both groups consume a comparable amount of informative content before they consider themselves to have learned enough about the topic.

User engagement was the matter of interest in Section 4.5, which provided both a ranking of all 150 topics and a ranking along the twelve users. Nine different features were incorporated for the computation of this ranking, and it turned out that the number of clicks, the overall reading time and the number of pastes are the most discriminating features for user engagement. Since we do not have assessments for the written essays, we were not able to check whether the user engagement correlates with the quality of the essays.

A new taxonomy for classification of different writing styles has been introduced in Section 4.6. This taxonomy contributes to the priorly classification scheme established by Potthast et al. [PHVS13a] in two points: First, by introducing an unambiguous, computable value for essay growth we overcome the issue that the original classification was based on human judges. Second, we argue that not only essay growth should be considered but also the regularity of paste events as this can show up trends that are not visible from essay growth. The resulting plot (cf. Figure 4.11) has not only proven the existence of different writing styles but also revealed that people pursue an individual working style, which they hardly can cast off. This finding is confirmed by further analyses, which are only marginally treated in Appendix A.

Finally, Section 4.7 examined different working phases during topic treatment. It was shown that the general trend exists that the number of queries, clicks and pastes decreases over time and the number of text-writing interactions increases. Yet, neither a clearly visible pattern could be identified nor was it possible to find characteristic differences along the users.

5.2 Future Work

The previous chapters have not only presented the raw analyses and their results but also how findings could contribute to current search engines in order to support users in exploratory search tasks. Some of the suggestions were, for example, to supply highly dedicated users with additional resources or to treat queries in a different manner depending on the working phase a user currently is in. While these approaches are rather long-term tasks, there are others that can be achieved in nearer future.

As already discussed in Section 4.3, for example, further effort can be put into development of a more sophisticated algorithm that determines how useful a clicked document is likely to be. If it was possible to achieve this task with a high accuracy, a search engine could particularly rely on terms of such useful documents for query expansion; or it could explore further documents that are linked within such a useful document, as we have shown that a quite large number of reference clicks appeared in course of click trails that contain at least one other reference document.

Another survey could examine why finders select their references more carefully as clickers do, which was discovered in Section 4.4. It would be also interesting to see if our ranking of user engagement correlates in some way with the quality of the produced essays, as assumed in Section 4.5. In this regard, an assessment of the essays is needed, which could be determined either automatically or manually. Further, we suggest to implement an algorithm which can identify an author just by the written essay, as left up for future work in Section 4.6. Last but not least, we made the conjecture in Section 4.7 that users despite of the ongoing time are not able to improve their precision in terms of needed queries and clicks to find a desired information. Further investigations are needed that examine whether this assumption really holds, and if so, why precision does not increase although searchers should have learned a lot about the topic and probably be able to better distinguish between useful and worthless sources.

Moreover, we suggest to make more use of the actual essay content in further studies. For example, key phrases could be extracted for each essay revision giving an indication for the content-related progress over all revisions. Assume that the final essay after 4,000 revisions contains 30 entities from the DBpedia[1] data set, and that revision 1,000 already contains 24 of them. So, the progress of essay content is already 80 % at that early point. Having extracted this information for each revision can reveal in which exact phases the user found particularly useful documents that helped to form large parts of the final essay.

[1] `http://dbpedia.org/`, Last accessed: August 30th, 2014

5.3 Discussion and Conclusion

Work on this thesis was as exploratory as the search tasks were for the users from our interaction logs. Certainly, valuable information can be found nearly everywhere and it is all about asking the right questions. Numerous analyses have been conducted during this thesis, far more than could have been presented here. Yet, as many of them seem promising, as many actually are not, or at least the data is too diverse to draw general inferences from them. When reading literature on exploratory search, one is often excited about the outcomes of a publication and, after having read it, a little bit disappointed as the outcomes often are so natural and intuitive that they seem not worth to even be mentioned. However, following the original goal of this thesis, namely to propose certain approaches that can improve search engines' assistance of users who are working on exploratory tasks, showed that these natural and intuitive findings are often not that easy to find and proof. Not to mention those outstanding and genuinely surprising outcomes that we have expected when starting work on this thesis.

So, to conclude this work, we consider the results that were achieved in this thesis, and which are briefly summarized in Section 5.1, to constitute one more little step on the long way down to an ideal search engine that perfectly fulfills the user's needs in any situation.

A Combination of Log Features

In Section 4.6 we combined two features, namely essay growth and paste regularity, in order to propose a new taxonomy for different writing styles. The plot in Figure 4.11 reveals that users tend to follow some behavioral patterns. Therefore, the data points often can be found in a particular area of the plot and form visual clusters. This leads to the assumption that further combinations of various features might reveal other findings of similar relevance. Therefore, we analyze each possible combination of the following features in a plot:

- Queries
- Distinct queries
- Median queries per session
- Distinct terms
- Median distinct terms per session
- Handled subtopics
- Clicks
- Distinct clicks
- Result clicks
- Trail clicks
- Percentage of trail clicks
- Ratio queries to clicks
- Reading time
- Ratio actual to expected reading time
- Essay length
- Revisions
- Pastes
- Used references
- Sessions
- Spanned days

- Spanned working days
- Working hours
- Time for text-writing

Since not all possible combinations of log features reveal unexpected findings, we present only the interesting ones in the following paragraphs.

Distinct Queries versus Clicks

Figure A.1, for example, shows the number of distinct queries on the x-axis and the number of clicks on the y-axis. As in Figure 4.11, most authors are located at only one particular area of the plot; the convex hulls for each user should emphasize this. Yet, what is even more interesting about Figure A.1 is the rudimentary presence of two arms extending from the lower-left area to the right or to the top of the plot, respectively.

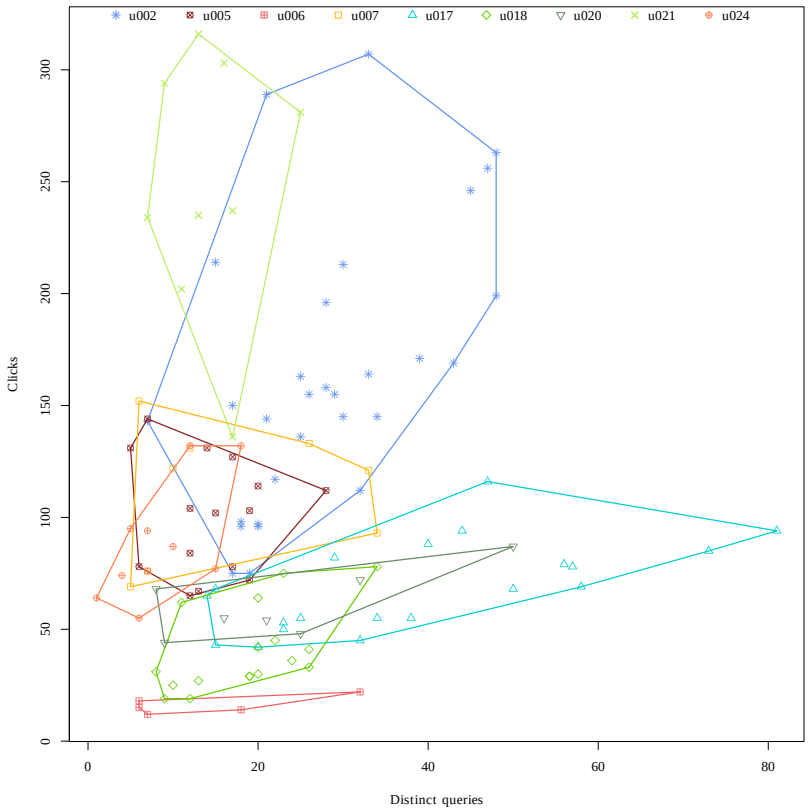


Figure A.1: Distinct queries vs. Clicks.

According to the location and shape of the convex hulls we can identify two different groups. The first comprises users 006, 017, 018 and 020, and the second group consists of users 002, 005, 007, 021 and 024. Retrospectively to Section 4.4, it turns out that these two groups correspond to the priorly identified groups of clickers and finders (except for users 002 and 020, which are erroneously interchanged).

Distinct Queries versus Percentage of Trail Clicks

Though a bit confusing because of the largely overlapping convex hulls, Figure A.2 again illustrates that almost all data points for one user are located at one specific area. Further, the plot perfectly meets our expectations about users who are rather trail clickers and those who are not: The more distinct queries are submitted, the less trail clicks occur – and the other way round.

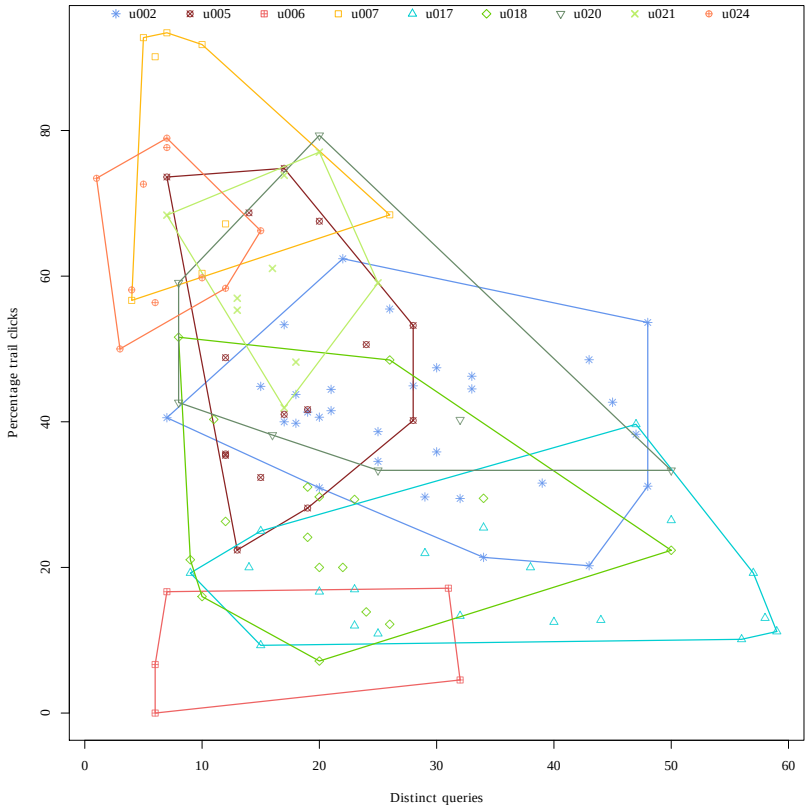


Figure A.2: Distinct queries vs. Percentage of trail clicks.

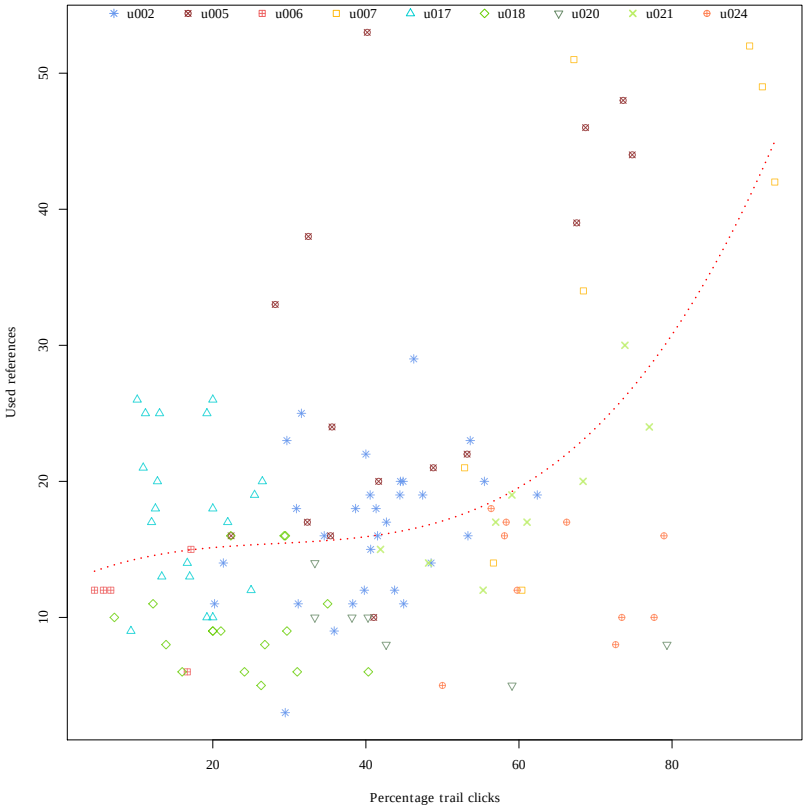


Figure A.3: Percentage of trail clicks vs. Used references.

Percentage of Trail Clicks versus Used References

Figure A.3 shows the percentage of trail clicks on the x-axis and the number of used references on the y-axis. In favor of less confusion, the clusters are not illustrated but a regression line has been added to the plot. This line reveals a slight trend that trail clickers use more references in their essays than result clickers do. However, note that the steep slope at the end of the line is only the consequence of very few data points and therefore should not be taken too serious.

Percentage of Trail Clicks versus Precision

Last but not least, we noticed an interesting pattern for the percentage of trail clicks on the x-axis and the precision on the y-axis, as can be seen in Figure A.4. Here, we define precision as the ratio of used references to distinctly clicked documents (i. e., documents that have been visited two or more times are only counted as one click).

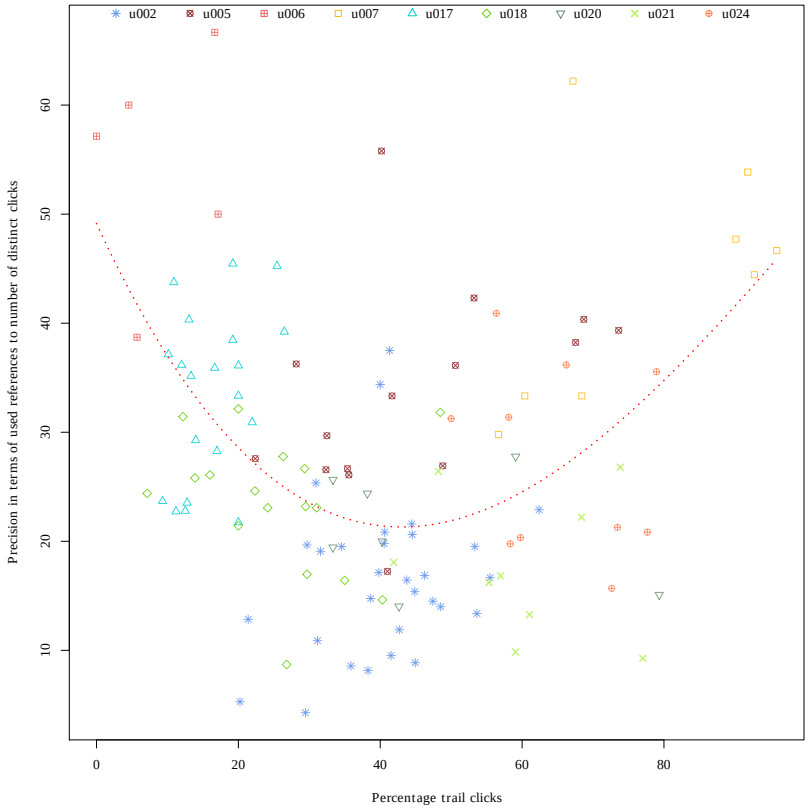


Figure A.4: Percentage of trail clicks vs. Precision.

Figure A.4 reveals that the precision is high for either result clickers and trail clickers but not for users that apply a mixed style. An explanation for this effect could not be found and is left up for future research.

List of Figures

List of Tables

Bibliography

[ABDR06] Eugene Agichtein, Eric Brill, Susan Dumais, and Robert Ragno. Learning User Interaction Models for Predicting Web Search Result Preferences. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 3–10, New York, NY, USA, 2006. ACM.

[Arg14] Jaime Arguello. Predicting Search Task Difficulty. In Maarten Rijke, Tom Kenter, Arjen P. de Vries, Cheng Xiang Zhai, Franciska Jong, Kira Radinsky, and Katja Hofmann, editors, *Advances in Information Retrieval*, volume 8416 of *Lecture Notes in Computer Science*, pages 88–99. Springer International Publishing, 2014.

[AWDB12] Eugene Agichtein, Ryen W. White, Susan T. Dumais, and Paul N. Bennet. Search, Interrupted: Understanding and Predicting Search Task Continuation. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 315–324, New York, NY, USA, 2012. ACM.

[BCC+09] Ranieri Baraglia, Fidel Cacheda, Victor Carneiro, Diego Fernandez, Vreixo Formoso, Raffaele Perego, and Fabrizio Silvestri. Search Shortcuts: A New Approach to the Recommendation of Queries. In *Proceedings of the 3rd ACM Conference on Recommender Systems*, RecSys '09, pages 77–84, New York, NY, USA, 2009. ACM.

[BWC+12] Paul N. Bennett, Ryen W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. Modeling the Impact of Short- and Long-term Behavior on Search Personalization. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 185–194, New York, NY, USA, 2012. ACM.

[CHYZ09] Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. The ClueWeb09 Data Set. Slides, 2009. `http://boston.lti.cs.cmu.edu/classes/11-742/S10-TREC/TREC-Nov19-09.pdf`, Last accessed: August 30th, 2014.

[CJP+09] Huanhuan Cao, Daxin Jiang, Jian Pei, Enhong Chen, and Hang Li. Towards Context-aware Search by Learning a Very Large Variable Length Hidden Markov Model from Search Logs. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 191–200, New York, NY, USA, 2009. ACM.

[Der92] Brenda Dervin. From the Mind's Eye of the User: The Sense-making Qualitative-quantitative Methodology. In Jack D. Glazier and Ronald R. Powell, editors, *Qualitative Research in Information Management*, volume 9, pages 61–84. Libraries Unlimited, Englewood, CO, 1992.

[DLDL65] M. De Leeuw and E. De Leeuw. *Read Better, Read Faster: A New Approach to Efficient Reading*. Pelican Original. Penguin Books, 1965.

[Eft96] Efthimis N. Efthimiadis. Query Expansion. *Annual Review of Information Science and Technology*, 31:121–187, 1996.

[EST+10] Yuka Egusa, Hitomi Saito, Masao Takaku, Hitoshi Terai, Makiko Miwa, and Noriko Kando. Using a Concept Map to Evaluate Exploratory Search. In *Proceedings of the 3rd Symposium on Information Interaction in Context*, IIiX '10, pages 175–184, New York, NY, USA, 2010. ACM.

[HGBS13] Matthias Hagen, Jakob Gomoll, Anna Beyer, and Benno Stein. From Search Session Detection to Search Mission Detection. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 85–92, Paris, France, 2013. ACM.

[KSGB12] Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. Age-of-Acquisition Ratings for 30,000 English Words. *Behavior Research Methods*, 44(4):978–990, 2012.

[Kur93] Martin Kurth. The Limits and Limitations of Transaction Log Analysis. *Library Hi Tech*, 11(2):98–104, 1993.

[Mar06] Gary Marchionini. Exploratory Search: From Finding to Understanding. *Communications of the ACM*, 49(4):41–46, 2006.

[OCDV12] Umut Ozertem, Olivier Chapelle, Pinar Donmez, and Emre Velipasaoglu. Learning to Suggest: A Machine Learning Framework for Ranking Query Suggestions. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 25–34, New York, NY, USA, 2012. ACM.

[PGH+12] Martin Potthast, Tim Gollub, Matthias Hagen, Jan Graßegger, Johannes Kiesel, Maximilian Michel, Arnd Oberländer, Martin Tippmann, Alberto Barrón-Cedeño, Parth Gupta, Paolo Rosso, and Benno Stein. Overview of the 4th International Competition on Plagiarism Detection. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *Working Notes Papers of the CLEF 2012 Evaluation Labs*, 2012.

[PHS+12] Martin Potthast, Matthias Hagen, Benno Stein, Jan Graßegger, Maximilian Michel, Martin Tippmann, and Clement Welsch. ChatNoir: A Search Engine for the ClueWeb09 Corpus. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 1004–1004, New York, NY, USA, 2012. ACM.

[PHVS13a] Martin Potthast, Matthias Hagen, Michael Völske, and Benno Stein. Crowdsourcing Interaction Logs to Understand Text Reuse from the Web. In Pascale Fung and Massimo Poesio, editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1212–1221. ACL, 2013.

[PHVS13b] Martin Potthast, Matthias Hagen, Michael Völske, and Benno Stein. Exploratory Search Missions for TREC Topics. In Max L. Wilson, Tony Russell-Rose, Birger Larsen, Preben Hansen, and Kristian Norling, editors, *3rd European Workshop on Human-Computer Interaction and Information Retrieval*, pages 11–14, 2013.

[PK82] E.J. Pedhazur and F.N. Kerlinger. *Multiple Regression in Behavioral Research: Explanation and Prediction*. Holt, Rinehart, and Winston, 1982.

[QF08] Yan Qu and George W. Furnas. Model-driven Formative Evaluation of Exploratory Search: A Study Under a Sensemaking Framework. *Information Processing and Management*, 44(2):534–555, 2008.

[RSPC93] Daniel M. Russell, Mark J. Stefik, Peter Pirolli, and Stuart K. Card. The Cost Structure of Sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*, pages 269–276. ACM, 1993.

[SGH11] Benno Stein, Tim Gollub, and Dennis Hoppe. Beyond Precision@10: Clustering the Long Tail of Web Search Results. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2141–2144, New York, NY, USA, 2011. ACM.

[Vak10] Pertti Vakkari. Exploratory Searching as Conceptual Exploration. *Proceedings of the 4th International Workshop on Human-Computer Interaction and Information Retrieval*, pages 24–27, 2010.

[VH12] Pertti Vakkari and Saila Huuskonen. Search Effort Degrades Search Output but Improves Task Outcome. *Journal of the American Society for Information Science and Technology*, 63(4):657–670, 2012.

[WKD+06] Ryen W. White, Bill Kules, Steven M. Drucker, et al. Introduction. In *Supporting Exploratory Search*, volume 49 of *Communications of the ACM*, pages 36–39. ACM, New York, NY, USA, 2006.

[WM07] Ryen W. White and Gary Marchionini. Examining the Effectiveness of Real-time Query Expansion. *Information Processing and Management*, 43(3):685–704, 2007.

[WMM08] Ryen W. White, Gary Marchionini, and Gheorghe Muresan. Editorial: Evaluating Exploratory Search Systems. *Information Processing and Management*, 44(2):433–436, 2008.

[WR09] Ryen W. White and Resa A. Roth. *Exploratory Search: Beyond the Query-Response Paradigm*, volume 1. Morgan & Claypool Publishers, 2009.

[XC00] Jinxi Xu and W. Bruce Croft. Improving the Effectiveness of Information Retrieval with Local Context Analysis. *ACM Transactions on Information Systems*, 18(1):79–112, 2000.