Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Computer Science and Media

Analyzing a Large Corpus of Crowdsourced Plagiarism

Master's Thesis

Michael Völske

First Supervisor: Prof. Dr. Benno Stein
Advisors: Dr. Martin Potthast,
Dr. Matthias Hagen,
Dr. Steven Burrows

Date of submission: March 20th, 2013

Abstract

This thesis investigates the Webis Text Reuse Corpus 2012 (Webis-TRC-12), a large dataset of crowdsourced plagiarism. In the process, it studies the categorization of text reuse strategies found in the corpus, and the patterns of search engine interaction of authors gathering material for a complex writing task. Its contributions include the processing and refinement of the data for future use, a framework for categorizing crowdsourced plagiarism, and a set of interactive tools that support the exploratory study of large corpora of this nature.

Declaration of Authorship

I hereby assert that this thesis is entirely my own work, employing only the reference media and sources.

Weimar, March 20th, 2013

. . . . . .
Michael Völske

Acknowledgements

This thesis would not have been possible without the help of many extraordinary people. First and foremost, I thank my advisors, Dr. Martin Potthast, Dr. Matthias Hagen, and Dr. Steven Burrows for innumerable helpful discussions. The countless hours of their time they donated to reviewing drafts of this thesis and pointing out new directions to explore made this work an enjoyable and productive learning experience.

I thank Prof. Dr. Benno Stein for taking this work under his wing as first supervisor, and for the many years spent working tirelessly to build the inspiring academic environment from which it grew.

I also thank the people who keep things running behind the scenes. Without the competent work of Nadin Glaser, Christin Gläser, and Dr. Berd Schalbe, my academic career would have been a much more stressful one.

I thank the people present at the National Polytechnic institute in Mexico in September 2012, including Enrique Flores Saez, Alexander Gelbukh, Ergina Kavallieratou, Grigori Sidorov, and Efstathios Stamatatos, for listening to a very early version of this research and providing helpful feedback.

I humbly thank my parents, Sibylle and Jörg Völske, for believing in me all these years, for teaching me to love science and language, and for instilling a thirst for knowledge that keeps taking me further down this rabbit hole.

Finally, I thank Jasmin Türker for her patience and support, and for keeping me sane throughout this massive undertaking.

Contents

1 Introduction **1**

2 Background and Related Work **3**
2.1 Evaluating Text Plagiarism Detectors 3
2.2 The Webis Text Reuse Corpus 2012 4
2.3 Previous Text Reuse Corpora 5
2.4 User Goals and Search Missions 7
2.5 Search Mission Corpora 9
2.6 Categorization of Plagiarism 11
2.7 Summary 13

3 Categorizing Crowdsourced Text Reuse **15**
3.1 A Plagiarism Spectrum for Corpus Documents 15
3.2 Measuring Interleaving and Paraphrasing 17
3.3 The Spectrum of Plagiarism in the Corpus 19
3.4 Document Change Over Time 21
3.5 Aggregated Editing Behavior by Author 23
3.6 Summary 25

4 Search Missions for Source Retrieval **26**
4.1 Properties of the Webis-TRC-12 Search Missions 26
4.2 Exploratory Search Missions in the Reference Corpus 29
4.3 Measuring the Similarity of Search Missions 29
4.4 Percentage of Queries Submitted Over Time 33
4.5 Correlating Querying and Editing Behavior 35
4.6 Summary 37

5 Corpus Construction and Statistics **38**
5.1 Corpus Generation Through Crowdsourcing 38
5.2 Document Statistics and Source Retrieval Models 44
5.3 Click Trails as Implicit Relevance Judgements 48
5.4 Interactive Corpus Tools 51
5.5 Summary 55

6 Conclusions **56**

Bibliography **59**

1 Introduction

Plagiarism is, in essence, the appropriation of another's ideas as one's own. Potthast (2011) summarizes the process of text plagiarism—text reuse without proper citation—as follows: the plagiarist copies passages from a source document into their own text, and usually modifies them in order to avoid detection. Plagiarizing from web sources thus involves three distinct steps: retrieving sources using a web search engine, extracting passages to plagiarize, and inserting them in a modified form into the final document.

An automatic system for detecting plagiarism from web sources would have to operate analogously: given a document suspected of plagiarism, it would have to retrieve potential sources from the web, identify the—potentially obfuscated—copied passages, and match each passage to the corresponding source. Comparing several different plagiarism detection systems with regards to their effectiveness requires (next to a set of appropriate performance measures) a standardized corpus of plagiarism, so that detectors' performance can be evaluated in a controlled setting.

The Webis Text Reuse Corpus 2012 (Webis-TRC-12) represents the latest effort to establish such a standard plagiarism dataset and forms the main basis for the work at hand. In order to collect the raw data for this corpus, Potthast et al. (2012a) employ 17 writers from an online crowdsourcing platform to write 297 plagiarized documents on 150 topics, each at least 5 000 words long.

What sets apart this dataset from previous corpora of this kind is the unprecedented level of detail in which it represents the process of plagiarizing from the web. All revisions an author makes to a document throughout the writing process are meticulously logged, providing insights into how authors copy, paste and rewrite plagiarized passages. In addition, plagiarists retrieve their sources using the ChatNoir search engine described by Potthast et al. (2012b). This custom search engine makes query logs and click trails available for further study.

Building upon this raw data, this thesis makes two major contributions: first, we take a detailed inventory of the dataset itself, refine it, and make it more accessible to subsequent analysis. Second, we apply the data to current research. While mainly intended for building a new testbed for automatic plagiarism detectors, we show that the data in the corpus can help test a variety of related hypotheses. Building on our post-processing and analysis of corpus data, we address the following research questions:

1. Is there evidence of distinct plagiarism strategies in the behavior of corpus authors, and if so, how do they relate to previous attempts at categorizing plagiarism?
2. How does a complex research task such as gathering sources for plagiarism reflect upon the author's interaction with the search engine, and how do the resulting query logs compare to other search engine interaction datasets?

We believe that the three fields of plagiarism, search, and paraphrasing research can benefit above all others from the data in Webis-TRC-12. Our first research question contributes to plagiarism analysis, also addressing paraphrasing aspects in the process. The second research question is mainly one of search engine log analysis, but has implications on plagiarism detection as well.

Thesis Organization

The remainder of this document is organized as follows: In Chapter 2, we give some background on how the data came into being, and survey the previous work relevant to our efforts. Chapters 3 and 4 document our investigations into the two research questions outlined above, as well as possible implications and interrelationships of our results. Chapter 5 examines additional data collected as part of the corpus and the implications of different retrieval models used during corpus creation. Finally, Chapter 6 summarizes and concludes this thesis, and points out possible avenues for future research.

2 Background and Related Work

This chapter introduces the necessary background for the remainder of this thesis. First of all, it describes the work undertaken by Potthast et al. (2012a) to prepare and collect the data for this thesis. Starting from the fundamental problem of constructing a reliable evaluation framework for text plagiarism detection in Section 2.1, we survey the corpus construction effort in Section 2.2, and present some key characteristics. In Section 2.3, we compare it to previous text reuse datasets.

The second half of the current chapter surveys the related work in fields relevant to our main research questions. In Section 2.4, we look at work concerned with the analysis of search engine query logs and search missions. The detailed search engine log collected as part of Webis-TRC-12 highlights the potential for new insights in these fields, especially when compared to the datasets used in previous efforts. The concept of a *search mission* is especially interesting in this regard, since the new query log constitutes the first publicly-available dataset where users' search missions are known a priori. We survey previous efforts at modeling search tasks, and existing search mission corpora in Section 2.5.

In addition to the above, Webis-TRC-12 presents a new opportunity to study the behavior of authors reusing text. In order to provide background to our efforts at categorizing plagiarists in Chapter 3, we survey previous work on this problem in Section 2.6.

2.1 Evaluating Text Plagiarism Detectors

As pointed out in Chapter 1, a standard dataset forms an integral part of an evaluation framework for automatic plagiarism detection. Potthast et al. (2010) survey 205 papers on automatic plagiarism detection in text and source code with respect to how the authors evaluate their detection methods. This analysis finds a lack of standardized evaluation resources for plagiarism detectors—only 20% (18%) of the papers on text (code) plagiarism detection use an existing corpus that would make their results reproducible. Subsequent efforts—such as the PAN corpora described by Potthast et al. (2010) and Potthast et al. (2011)—have begun to address this issue.

Potthast (2011, pg. 75) points out three fundamental approaches to acquiring cases of plagiarism for use in a corpus. Collecting cases of *real plagiarism* is inadvisable

for legal and ethical reasons, as well as the fact that a corpus comprised of such will necessarily be biased towards more easily detected instances. On the other end of the spectrum is *artificial plagiarism*, where documents containing foreign material are generated algorithmically. *Simulated plagiarism* forms a middle ground, in that a paid or volunteer worker—instructed to plagiarize in as realistic a setting as possible—"goes through the motions" of producing a plagiarized text. This is the approach behind the data collection for Webis-TRC-12.

2.2 The Webis Text Reuse Corpus 2012

Potthast et al. (2012a) organized the data collection effort for Webis-TRC-12. Their work includes formulating the task descriptions for the different topics in the corpus, hiring and managing crowdsourcing workers, as well as creating document editing and search engine interfaces for them to use in their task. All told, the corpus is comprised of 297 documents written by 27 individuals, all of them experienced in writing English text.

While some authors are volunteers recruited from university staff, most of the documents were authored by professional writers hired on the online crowdsourcing platform oDesk.[1] The use of crowdsourcing to drive the construction of evaluation resources has become commonplace in fields as diverse as search engine evaluation (Carvalho et al., 2011) and vandalism detection in Wikipedia (Potthast, 2010). While Amazon's Mechanical Turk remains the platform of choice in most of these cases, oDesk offers several advantages when dealing with complex tasks such as the Webis-TRC-12 corpus construction effort. Its detailed worker profiles and employment histories allow the selection of experienced writers, while worker tracking systems and reputation scores reduce the likelihood of workers submitting fake results. For privacy reasons, individual authors will only be identified by alphanumerical aliases—such as A001—in the remainder of this thesis.

A distinguishing feature of Webis-TRC-12 is the fact that it builds upon other, well-established information retrieval data sets. Potthast et al. (2012a) defined 150 topics for the different documents in the corpus. Since the Text Retrieval Conference (TREC) has famously established a number of standard information retrieval topics, Potthast et al. (2012a) used the 150 topics from the TREC Web tracks 2009–2011[2] as a starting point. The sources for all the plagiarized documents in the corpus are taken from the English language subset of the ClueWeb09 corpus.[3] ClueWeb09—a large-scale web crawl—comprises over 1 billion pages in total. Taking the sources of plagiarism from a set of documents of this magnitude makes the resulting corpus useful for evaluating web-scale source retrieval techniques in a realistic setting. Each

[1] `https://www.odesk.com/` (last accessed March 2013)
[2] `http://plg.uwaterloo.ca/~trecweb/` (last accessed March 2013)
[3] `http://lemurproject.org/clueweb09.php` (last accessed March 2013)

of the TREC topics defines an information need, as well as a set of ClueWeb09 documents judged as relevant (or not) by TREC assessors. While Potthast et al. (2012a) rephrase the topics to motivate writing an essay, the underlying information need remains the same. This provides a starting point for evaluating the query logs in Webis-TRC-12.

The corpus authors wrote their essays in two batches; in the first, authors used only the top ranked ClueWeb09 documents—as determined from the TREC relevance judgements—as potential sources for plagiarism. In the second batch, authors were instead given access to the ChatNoir search engine described by Potthast et al. (2012b), and instructed to find their own sources among the full set of ClueWeb09 documents. Each topic was used once per batch (excluding three topics in Batch 1 for which no relevant ClueWeb09 sources were available), and no author was assigned the same topic twice. As explained by Potthast et al. (2012a), this helps control for the effects of the different retrieval models on the selection of source documents. Unless noted otherwise, the remainder of this thesis concerns itself with the subset of documents in Batch 2 when discussing authors' interaction with the search engine, and with the entire dataset in all other cases.

In order to better judge the novelty of the Webis-TRC-12 dataset, previous efforts at text reuse datasets are of interest.

2.3 Previous Text Reuse Corpora

Potthast et al. (2010) note a general lack of evaluation resources for text reuse detection and introduce the PAN-PC series of corpora to help fill this gap. Table 2.1 shows basic statistics for the PAN-PC-10 dataset, as well as the subsequent PAN-PC-11 dataset introduced by Potthast et al. (2011). These ancestors of the Webis-TRC-12 were constructed for the PAN plagiarism detection competitions and contain mostly algorithmically generated cases of plagiarism, produced by randomly selecting passages from source documents and inserting them into another host document. The PAN-PC corpora also model the obfuscation of plagiarized passages. In most cases, this is done algorithmically—by randomly inserting, replacing, deleting or shuffling the words in the plagiarized passage, or by inserting automatically-translated foreign language text.

While these automatic transformations enable the fast generation of a large repository of artificially-plagiarized documents, they generally do not preserve the semantics of the modified passage. To counter this issue, a small subset of plagiarism cases were rewritten manually by crowdsourcing workers recruited on Amazon's Mechanical Turk. However, as pointed out in Potthast et al. (2011), this does not solve the problem of missing topic overlap between host document and inserted passage. Since the source documents for plagiarized passages are chosen at random, they probably introduce words into the text that would not occur otherwise. When evaluating

Table 2.1: Characteristics of two previous PAN corpora.

	PAN-PC-10	**PAN-PC-11**
Documents	**27 073**	**26 939**
- source documents	13 536	13 470
- with plagiarism	6 768	6 735
- without plagiarism	6 768	6 735
Plagiarism Cases	**68 558**	**61 064**
- unobfuscated	27 423	10 992
- paraphrased (automatic)	27 423	38 470
- paraphrased (manual)	4 113	4 885
- translated	9 598	6 717

word-based retrieval models (e.g. topic drift analysis), the corpora may thus not present a realistic evaluation scenario.

As Table 2.1 shows, the PAN-PC corpora are much larger in magnitude than the Webis Text Reuse Corpus 2012. However, since the latter corpus is constructed manually in its entirety, it does not suffer from any of the drawbacks of automatically constructed corpora mentioned above.

Burrows et al. (2012) tackle the related problem of constructing a corpus of passage-level paraphrasing samples. In their survey of existing paraphrasing corpora, they note that while several collections of sentence-level paraphrases exist, corpora of larger text units are rare. However, to a number of paraphrasing related tasks, including plagiarism detection, sentence-level corpora are of little use. To address this issue, Burrows et al. (2012) introduce the PAN-CPC-11 corpus comprising 7 859 positive and negative samples of passage-level paraphrasing. To acquire paraphrases, they too employ crowdsourcing via Amazon Mechanical Turk.

Aside from the PAN corpora, there are only very few evaluation datasets for plagiarism and text reuse. One of them is the METER corpus of journalistic text reuse published by Clough et al. (2002). This corpus comprises 445 cases of text reuse among 1 716 news articles, taken from real-world occurrences of text reuse in a journalistic context. The text reuse samples in the METER corpus vary in length, from one-sentence summaries to longer reports of several hundred words. Due to the different scope, it is not well suited as an evaluation resource for plagiarism detection.

The Clough09 corpus of plagiarized short answers, published by Clough and Stevenson (2011), constitutes another text reuse dataset. For this dataset, a number of volunteer students wrote a set of 57 short answers to five different computer science questions, while reusing text from predetermined Wikipedia articles. At only a few hundred words each, the documents in this corpus are rather short; this limits the

corpus' usefulness as an evaluation resource. Moreover, since the Wikipedia articles were given up front, the Clough09 corpus does not support a study of source retrieval, as is possible using the search engine logs that are part of Webis-TRC-12.

In order to be better able to organize and catalogue the information in the search engine logs, we next survey efforts at modeling and understanding user goals in information retrieval.

2.4 User Goals and Search Missions

The information in Webis-TRC-12 includes a detailed search engine interaction log. As such, we are interested in comparing the Webis-TRC-12 query logs to prior discussions of the subject of user interaction with search engines. On the one hand, this allows us to establish a common terminology with which to describe our data set. On the other hand, we can determine if the behavior of our authors shows characteristic patterns that can be found in other available datasets.

In order to describe the query log, this thesis adopts the terminology of Jones and Klinkner (2008), who define a *search mission* as a related set of overarching information needs that direct a user's querying behavior. The queries that pursue a given search mission may be split across several *search sessions*. A search session is defined as a period of activity, during which the user may pursue just one, or several interleaved search missions. Jones and Klinkner (2008) study in particular the situation where multiple distinct search missions are interleaved throughout a given time frame. By contrast, in our query logs, all queries submitted for a given topic correspond to the same search mission—which in turn is driven by the topic that the user is writing about.

For the purposes of our investigation, we assume that all of our authors pursue search missions from the same category, since their search engine use supports the same type of task—that of writing a long article using web sources. In order to arrive at a consistent terminology for this type of search task, we investigate two previous systems of user task classification: information gathering and exploratory search.

Information Gathering

Past research has established different frameworks for understanding the user intent behind search engine queries, as well as to automatically infer it from search behavior. Broder (2002), Kellar et al. (2007) and Sellen et al. (2002) propose sets of broad categories to classify the user's information need. Table 2.2 contrasts their respective classification systems.

Table 2.2: Three frameworks for understanding the user intent behind search tasks.

Broder (2002)	**Kellar et al. (2007)**	**Sellen et al. (2002)**
Navigational	Browsing	Browsing
Informational	Fact Finding Information Gathering	Finding Information Gathering
Transactional	Transactions	Transacting Communicating Housekeeping

Broder (2002) distinguishes *navigational*, *transactional* and *informational* search engine queries. In this simple taxonomy, only the latter type of query is driven by an actual search for information. With the former two, the user is trying to reach some particular web site, or perform some web-based activity (Broder (2002) names shopping or downloading files as examples), respectively.

Kellar et al. (2007) and Sellen et al. (2002) further partition the category of informational queries into two distinct types. *(Fact) finding* is concerned with retrieving some specific, atomic piece of information, such as a name, a phone number, or an image. Sellen et al. (2002) distinguish additional types of transactional user intents concerned with online discussion and maintenance of web resources.

The search missions in Webis-TRC-12 would be sorted into the broad category of informational queries in the terminology of Broder (2002), whereas Kellar et al. (2007) and Sellen et al. (2002) would refer to them more specifically as information gathering.

Exploratory Search

Marchionini (2006) presents a very different hierarchy of search goals and supporting strategies, which is summarized in Table 2.3. This taxonomy classifies basic information needs ("tasks") by the kinds of search strategies ("activities") that may be used to accomplish them. In stark contrast to Broder (2002) and related taxonomies, the information needs behind fact finding, navigational, and transactional tasks all fall into the *lookup* category, since the same type of search strategy applies: a small number of carefully phrased queries is used to retrieve some specific, atomic piece of information that the user is looking for. This category best fits the classical query-response paradigm of information retrieval.

By contrast, users pursuing tasks in the *learn* and *investigate* categories are not aware beforehand of the full extent of the knowledge that they are seeking. As described by White et al. (2006), they need to take a much more ad-hoc approach: initial, tentative queries aimed at gaining a broad understanding of the subject are

Table 2.3: Taxonomy of search tasks and activities described by Marchionini (2006).

Activity	Task	
Lookup	Fact Finding Known Item Finding Navigation Transaction	
Learn	Knowledge Acquisition Interpretation Comparison Socialization	**Exploratory Search**
Investigate	Analysis Evaluation Discovery Planning	

gradually refined and expanded as the user's knowledge increases. Especially in the *investigate* category, this may lead to the pursuit of new domains of knowledge altogether.

The categories of *lookup*, *learn* and *investigate* are not sharply demarcated; rather, they form a continuum ordered by increasing depth of knowledge that the user is seeking. As Nolan (2008) points out, all three categories may apply throughout the pursuit of a single search mission. As the overlap between learning and investigative search behavior is especially pronounced, Marchionini (2006) aggregates them under the label of *exploratory search*.

Based on the preceding survey, we are presented with several options to name the search missions in Webis-TRC-12. In the remainder of this thesis, we adopt *exploratory search*, being the term that most broadly fits the set of tasks our authors needed to perform in order to consolidate several previously unknown sources into a coherent result.

In order to get the most use out of the Webis-TRC-12 search data, we will need to compare it to a reference dataset. To that end, we next survey previous efforts at modeling and predicting user behavior, in search of a suitable corpus for comparison.

2.5 Search Mission Corpora

Given that the query log from Webis-TRC-12 seems to represent a novel quality of search mission dataset, we are interested in comparing it to others. As a first step

towards such an analysis, this thesis explores the degree to which the search missions in Webis-TRC-12 are qualitatively different from the ones in other corpora. To this end, previous work on modeling and classifying search tasks is of interest.

Predicting Task Continuation

Both Kotov et al. (2011) and Agichtein et al. (2012) describe efforts at automatically inferring the class of information need at the bottom of a given search mission. They argue that some search tasks are especially complex, and as such are more likely than others to span multiple search sessions. Their research aims at predicting whether or not a given search mission will be resumed in the future, based on past characteristics of the query log for that mission. This has immediate applications to search engine design: once an information retrieval system detects a search mission that is likely to be resumed, it can record contextual information to help support future search sessions.

The work of Agichtein et al. (2012) work uses that of Kotov et al. (2011) as a baseline, and the feature sets used to predict task continuations overlap. We use the combined feature sets of both as a starting point for our own investigation in Chapter 4. However, only a small subset is actually part of our analysis, due to the exploratory nature of our research, as well as most of them being specific to the problem of predicting task continuation.

While both Kotov et al. (2011) and Agichtein et al. (2012) report on experiments with search log datasets, none of the corpora used in their work is available to the public.

The Webis-SMC-12 dataset

Hagen et al. (2013) present the Webis-SMC-12 search mission detection corpus comprising almost 10 000 queries of 127 users sampled from the AOL query log (Pass et al., 2006). Hagen et al. (2012) manually annotate each query as belonging to one of roughly 1 200 distinct search missions. The subset of the AOL log used to construct the corpus is based on a sample published by Gayo-Avello (2009). Since the original sample was aimed at retaining representative querying behavior, we can assume that Webis-SMC-12 retains this property after the annotation process. The resulting large dataset of search missions presents an interesting opportunity to compare the data in the Webis-TRC-12 query log to user interaction with a real-world web search engine. In Chapter 4, we investigate whether similar patterns of interaction can be found in both corpora.

The only other public search mission corpus is published by Lucchese et al. (2011). This dataset—also sampled from the AOL query log—consists of 1 424 queries from

13 users. Due to the much larger magnitude, we use Webis-SMC-12 for our experiments.

Aside from user search behavior, another major subject of interest in our analysis of the data is to catalogue the behavior of corpus authors while plagiarizing. To that end, we next survey a set of previous efforts at categorizing plagiarism.

2.6 Categorization of Plagiarism

In order to help make sense of the wealth of data in Webis-TRC-12, a taxonomy of the types of plagiarism it contains is a desirable goal. The corpus is the first public dataset of its depth, and the analysis presented in this thesis is only the first step in making it accessible.

Providers of commercial plagiarism detection systems, such as TurnItIn[4] already have access to large repositories of real-world plagiarism cases. Not least for ethical and business reasons however, their knowledge will never be fully public. Studies such as Turnitin (2012) provide a glimpse into their experiences. Based on a survey of 879 education professionals, the authors present a set of ten categories for plagiarism cases occurring in a classroom setting. They rank these plagiarism types both by how often they occur—in the experience of the study respondents—as well as by how "problematic" they are. The latter metric aims to take the assumed intent behind the plagiarism case into account. Since the plagiarism cases studied by Turnitin (2012) are collected from the work of students, they may not be attempts to actively deceive, but in fact just examples of poor scientific work due to lack of experience. Table 2.4 summarizes their findings.

A key insight regarding the TurnItIn categories is the emphasis on the educational aspect. For instance, while "Mashup" and "Remix" are essentially the same from an academic honesty standpoint, the additional work of paraphrasing that students have to perform to produce the latter is counted in their favor: "Remix" is ranked as much less severe. Of course, the educational point of view is irrelevant when categorizing plagiarism in our crowdsourced corpus.

The VroniPlag Wiki—a collaborative effort at investigating plagiarism in German doctoral theses—presents a set of categories distinct from the TurnItIn ones. VroniPlag (2012) lists the categories of plagiarism cases used in their annotation work. Since VroniPlag deals with the search for plagiarism in long academic texts, their categories deal with classifying the plagiarized text fragments rather than the suspicious document as a whole. Table 2.5 summarizes their main categories, translated from German.

[4] `http://turnitin.com/` (last accessed January 2013)

Table 2.4: The types of plagiarism reported by Turnitin (2012).

Type	**Rank**	
	frequency	severity
Clone Exact copy of another author's work.	1	1
Mashup A mix of material copied verbatim from several sources.	2	3
Ctrl-C Significant portions of text copied from a single source.	3	2
Remix Paraphrasing from several sources and making the content fit together seamlessly.	4	9
Recycle Self-plagiarism	5	5
Re-Tweet Proper citation, but closely follows a single source.	6	10
Find-Replace Near copy of a single source, with key phrases changed.	7	7
Aggregator Proper citation, but (almost) no original work.	8	4
404 Error Citations to non-existent or inaccurate information about sources.	9	6
Hybrid Combining properly cited sources with plagiarism in one paper.	10	8

Table 2.5: Categories of plagiarism described by VroniPlag (2012), and the corresponding Turnitin (2012) types that best match each.

VroniPlag	TurnItIn
Outright Plagiarism Text of the source is copied unmodified, without citation.	Clone, Ctrl-C, Recycle
Obfuscation Lifted text is modified slightly; no citation.	Find-Replace, Remix
Pawn Sacrifice Copied text does contain a footnote to the original. However, the suspicious document either fails to indicate the fact that text has been borrowed verbatim, or the full extent of the appropriation is obfuscated (e.g. by failing to indicate subsequent passages borrowed from the same source).	(Hybrid)
Translation Plagiarism Literal translation of a foreign-language text without proper citation.	N/A

Despite the different approach, there are obvious connections between the VroniPlag and the TurnItIn categories. The second column in Table 2.5 shows which of the TurnItIn types best match the four VroniPlag categories. The link between "Pawn Sacrifice" and "Hybrid" is only tenuous, and Turnitin (2012) does not consider translation plagiarism.

For the analysis of the plagiarism in Webis-TRC-12, most of the categories proposed by both sources don't make much sense. Since our authors were explicitly instructed to reuse text, the categories that include (partial) citations cannot apply. Translation plagiarism cannot apply either, since all documents and sources are in English. However, it may be of interest to see where each document falls in a continuum delineated by several of the TurnItIn categories.

2.7 Summary

This chapter introduced the necessary background for the work presented in the following chapters. Starting from the fundamental problem of evaluating text plagiarism detection, we gave a brief history of the Webis-TRC-12 dataset and how it came into being. We followed up with a survey of alternative text reuse datasets, which indicates that Webis-TRC-12 is indeed of a novel quality. Next, we introduced some related work on the study of search missions to help establish a consistent terminology, and to frame our own investigations in Chapter 4.

In the process, we introduced the Webis-SMC-12 search mission dataset, which is instrumental to our understanding and analysis of the query logs in Webis-TRC-12. In Chapter 4, we use Webis-SMC-12 as a reference corpus against which we compare the Webis-TRC-12 search missions.

Finally, we presented an analysis of previous efforts to categorize plagiarism. We found that the categories used to classify and analyze plagiarism in the real world may not be well-suited to the study of a crowdsourced dataset of simulated plagiarism. In Chapter 3, we derive a framework that may be more appropriate to our scenario, and investigate how well it fits the data.

3 Categorizing Crowdsourced Text Reuse

In this chapter, we examine how the text reuse behavior exhibited by the corpus authors can be mapped to the categories of plagiarism found in the literature. In Section 2.6 we presented two taxonomies of plagiarism in common use. We also explained that most of the categories covered do not apply to a corpus of simulated plagiarism, since the corresponding aspects of real-world academic dishonesty are not modeled by such a corpus.

However, a subset of the categories described in the related work is still useful for our purposes, as they map out a space of possibilities for the properties of the corpus documents. In Section 3.1, we derive a framework from some of the plagiarism categories found in the literature that map to the properties of our corpus documents. We then describe a simple hypothesis that can shed some light on the question of whether our framework is useful in distinguishing the kinds of plagiarism found in the Webis-TRC-12 dataset.

In Section 3.2 we prepare a simple experiment to test this hypothesis, and introduce measures to quantify the dimension of our plagiarism spectrum. Our results, described in Section 3.3, show that our measures do achieve the desired mapping of corpus documents into the space of plagiarism categories. We find evidence that the properties of a document may identify its author's modus operandi.

Section 3.4 examines an additional dimension of our corpus that we may explore beyond what is found in the literature: the way plagiarized texts evolve over time. In Section 3.5, we investigate how this property correlates with the author dimension.

3.1 A Plagiarism Spectrum for Corpus Documents

Based on the related work discussed in Chapter 2, we find two key aspects of the plagiarism categories applicable to corpus documents: the degree of *paraphrasing* applied to passages from the plagiarized sources, as well as by the degree to which these passages are *interleaved*.

Figure 3.1 shows our attempt to organize these properties into a framework for characterizing corpus documents. A document with a low degree of paraphrasing will contain mostly verbatim copies, whereas in a document with a high degree of paraphrasing, most source passages will have been rewritten. Highly interleaved

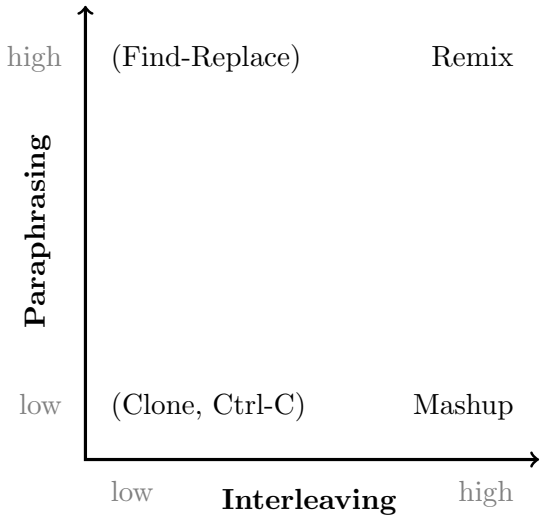


Figure 3.1: Proposed spectrum for the plagiarism in corpus documents. Categories in brackets are unlikely to be found in Webis-TRC-12, but are useful to demarcate the spectrum of possibilities.

documents will consist of many short passages alternating between different sources; documents with little interleaving will mostly consist of long blocks of contiguous text lifted from the same source, and perhaps fewer distinct sources in total.

Figure 3.1 also shows how some of the Turnitin (2012) categories from Table 2.4 would fall into this spectrum. Especially the "Find-Replace" and the "Clone" and "Ctrl-C" denote extremes that probably won't be attained by the documents in the corpus: it seems unlikely that many of them will draw from only a single source.

In order for our plagiarism spectrum to make sense, we must show that the documents in the corpus can be meaningfully distinguished along the two dimensions proposed in Figure 3.1. Since document properties measured using numerical features will no doubt be noisy, we do not expect the classification to be as clear-cut as the one proposed by Turnitin (2012). Instead, a given document will fall somewhere along the spectrum described by each axis of Figure 3.1.

Apart from measuring properties of the plagiarized text in corpus documents however, we can also examine the author dimension. The Webis-TRC-12 authors were instructed to plagiarize in such a way as to avoid detection, but the amount of obfuscation necessary to achieve this was left up to the individual to decide. Thus, if interleaving and paraphrasing are useful measures of crowdsourced plagiarism, they will reflect the individual working habits of the authors.

Based on these deliberations, we formulate the following hypotheses:

Hypothesis 3.1. *The degree to which corpus documents have been interleaved and paraphrased can be mapped to the spectrum outlined in Figure 3.1.*

Hypothesis 3.2. *Documents written by the same author tend to be similar with respect to these measures.*

Before we can test these hypotheses, we first need to narrow down the subject of our investigation, and define a set of empirical measures to quantify interleaving and paraphrasing.

3.2 Measuring Interleaving and Paraphrasing

While writing the documents in Webis-TRC-12, authors highlighted sections to indicate the source document in the ClueWeb09 corpus for the corresponding reused text. For any given sequence of words in a corpus document, we know whether and from where it was plagiarized, or whether it contains original work. Thus, we can choose an appropriate level of granularity at which to measure corpus documents, using information about the contained words and their source.

Passages of Plagiarized Documents

With respect to paraphrasing, we want to measure how much the text lifted from a given source has been modified. In terms of interleaving, we are interested in the way parts from different documents mesh together to form the final result. To meet both of these needs, we define the *passage* as the suitable unit of granularity at which we want to study the corpus. For the purpose of this experiment, a passage (or, "chunk") c of a document d refers to any *contiguous* block of text plagiarized from the same source document S. Thus, a document is a sequence of passages, $d = \{c_1, \cdots, c_n\}$ with the following properties:

1. Any pair of consecutive passages c_i and c_{i+1} have different sources.
2. For any passage c_i there may or may not exist a passage $c_j, j \notin \{i-1, i, i+1\}$ lifted from the same source.

For simplicity, we ignore any markup in the corpus document that may indicate a further subdivision of a passage into parts (perhaps lifted from different sections of the same source document). Instead, for all words in the document, we assume that two subsequent words from the same source will always belong to the same passage.

We extract the sequence of passages from each document to compute the measures of n-gram source similarity and passages per source, defined below.

N-gram Source Similarity

In order to quantify paraphrasing, we measure the degree of similarity of the text in the plagiarized passages to the text in the source document. To this end, we employ a one-way n-gram similarity.

For a given value of n, the n-grams in a passage of text form the set of all n-tuples of consecutive words. Thus, the sentence *"Writing is easy."* contains the 1-grams $\{(writing), (is), (easy)\}$, the 2-grams $\{(writing, is), (is, easy)\}$, and the single 3-gram $(writing, is, easy)$.

To compute an n-gram similarity for a pair of texts and a value of n, we first generate the sets of n-grams for both, and then compute the size of the intersection, i.e., the number of n-grams contained in both sets. Typically, the n-gram similarity is expressed as a fraction of the total number of n-grams in the union of both n-gram sets. In our case however, we are comparing a plagiarized text passage to a source document, where one of the texts is probably much shorter than the other.

What's more, we are really only interested in the n-grams of the plagiarized passage; additional content in the source document that is not used in a given passage should not be considered. We formulate the following one-way similarity function φ_n for the set of n-grams N_c in the plagiarized passage, and the set N_s of n-grams in the source:

$$\varphi_n(N_c, N_s) := \frac{|N_c \cap N_s|}{|N_c|}$$

Thus, the one-way similarity for a given passage is defined as the fraction of n-grams in the passage that also occur in the source. The function's values are in the range $[0, 1]$.

To determine an appropriate value of n, we consider the following constraints:

1. The similarity function should distinguish paraphrased passages from text copied verbatim. For instance, when $n = 1$, the similarity will remain the same under paraphrasing operations such as changing word order.
2. To get a good overview of the range of paraphrasing in the corpus, the number of similar and dissimilar passages in the entire corpus should be approximately equal.
3. It should be possible to compute the similarity for short passages of text. If n is larger than the number of words in a passage, the similarity cannot be computed.

The former two of these constraints ensure that the measure usefully divides the corpus documents for our exploratory study, and the latter enables us to study as large a portion of the data as possible.

We identify $n = 5$ as the best choice for our needs. Of all 3 905 passages collected across the 297 corpus documents, 99% contain at least 5 words. Out of those, 51% have a φ_5-value less than or equal to 0.5.

Since we use a passage-level similarity function, we get one similarity value for each passage in a corpus document. As a simple measure of the typical degree of paraphrasing found in a given document, we pick the median passage similarity.

Passages Per Source

For a measure of interleaving, we consider both the number of passages and the number of sources in a corpus document. In keeping with Figure 3.1, we want a measure that increases with the degree of interleaving. Documents with a higher degree of interleaving will contain more passages, but so will longer documents with more sources in total. Thus we define the function *pps* on the set of passages C and the set of sources S of a given document to be simply the ratio of the number of passages to the number of sources:

$$pps(C, S) := \frac{|C|}{|S|}$$

This function assumes its minimal value of 1 when a corpus document contains exactly one passage from each source. Its value increases as content from different ClueWeb09 documents is interleaved, resulting in more passages per source.

Having defined the set of objects we want to investigate, and appropriate measures to quantify their properties, we proceed to evaluate the hypotheses outlined in Section 3.1.

3.3 The Spectrum of Plagiarism in the Corpus

Figure 3.2 shows a scatter plot of the interleaving and paraphrasing measures for each document in the Webis-TRC-12 corpus. The x-axis shows the degree of interleaving as measured in passages per source. This measure follows a long tailed distribution, with a maximum of 11.5, a 90th percentile of 5.2, and a median of only 1.9. Hence, we choose a logarithmic x-axis with base 2 to provide a better overview of the spread of the data.

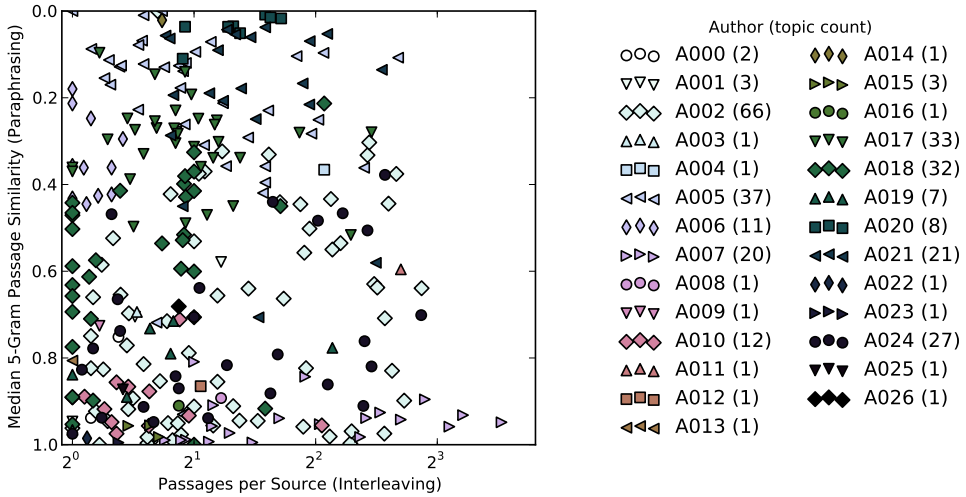


Figure 3.2: Spectrum of interleaving and paraphrasing in the corpus. Each data point corresponds to one corpus document. Note the inverted y-axis to match Figure 3.1. The x-axis is logarithmic.

The degree of paraphrasing as measured by median 5-gram similarity is shown on the y-axis. Since a higher degree of paraphrasing actually correlates with a *lower* n-gram similarity, we invert the y-axis to better correspond to Figure 3.1.

The distribution of document properties gives some evidence in favor of Hypothesis 3.1. However, as is evident from Figure 3.2, not the entire spectrum of values is actually covered. Documents with a low degree of interleaving occur with high as well as low paraphrasing, whereas the six most interleaved documents all have a below-average degree of paraphrasing. This might be due to the fact that the passage-to-source ratio does not quantify the amount of interleaving in the document very well.

It might however be evidence of an actual inverse correlation between the amount of interleaving effort and the amount of paraphrasing effort that authors tend to bring to the task. Further experiments may be able to shed light on this question. For instance, manually annotating some corpus documents regarding the degree of interleaving and paraphrasing perceived by a human evaluator may provide a standard against which our measures can be evaluated.

The evidence regarding Hypothesis 3.2 is more strongly in favor of our supposition that authors will be distinguishable by their paraphrasing and interleaving behavior. Figure 3.2 shows a strong tendency for documents written by the same author to cluster together. For instance, Authors A007 and A010 both tend to paraphrase very little compared to other authors. However, the former author employs much more interleaving than the latter in the majority of cases. By contrast, all documents

by Author A006 are on the lower end of the interleaving spectrum, but tend to be above average in terms of paraphrasing.

Not all authors are clustered as tightly as these examples. For instance, the many documents written by Author A024 spread across the entire interleaving spectrum, but never stray far above an average paraphrasing score. There is much potential for further investigation; for example, a simple classification experiment may help quantify how well these and other measures separate classes when used as features for authorship attribution of corpus documents.

So far, we have only investigated properties of corpus documents that can also be found in cases of real plagiarism. By virtue of recording all changes made to documents over time, Webis-TRC-12 contains at least one additional dimension that we have yet to explore. The remainder of this chapter addresses the time dimension of corpus documents.

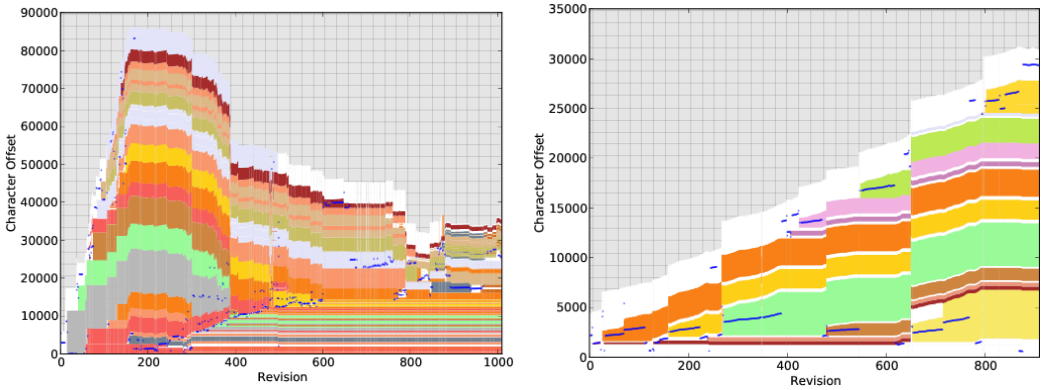
3.4 Document Change Over Time

Given the wealth of data available, we can compare the plagiarized documents in Webis-TRC-12 on a deeper level than simply looking at the final state of the plagiarized text. In order to be able to compare documents' revision histories, we develop a visualization that condenses a document's entire history into a single graph. The edit history visualization is based on a similar idea to the one presented by Viégas et al. (2004). While Viégas et al. (2004) visualize the contributions of different *authors* to a Wikipedia article over time, our main goal is to show how a corpus document is composed out of text passages lifted from different *sources* throughout the writing process.

Figure 3.3 shows some examples. The revision number is mapped to the x-axis; the y-axis represents the character position within the text. The color of a given pixel is based on the source from which the plagiarist lifted the corresponding portion of the document—white regions denote text marked as original content. Thus, any given vertical column in the graph illustrates the composition of the document at a specific moment in its editing history.

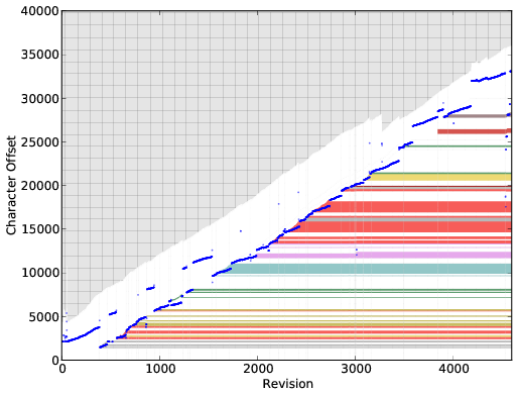
This bird's-eye view of a text's editing history provides some interesting insights into how different people approach the task of writing a plagiarized text. While some prefer to first paste in lots of copied text that is then condensed into a shorter final document (Figures 3.3a, 3.3b, and 3.3f), other authors work in a more linear fashion, editing copied content right after pasting it in (Figures 3.3c and 3.3b). We refer to the former as *boil-down*, and the latter of *build-up* editing behavior.

A cursory visual analysis of plots like the ones in Figure 3.3 for all corpus documents reveals some additional statistics about the composition of the corpus. Out of the 297 documents in the corpus, 166 (39%) are in the build-up category, whereas 104 (35%)

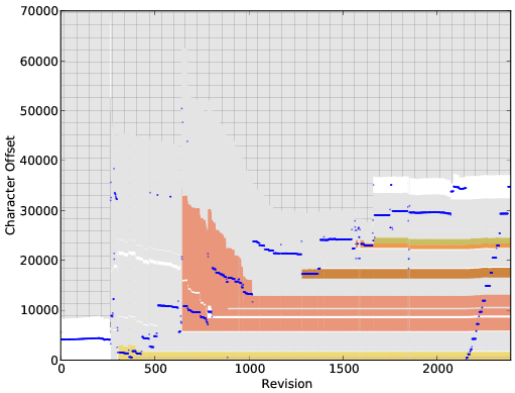


(a) Topic 88, Batch 1

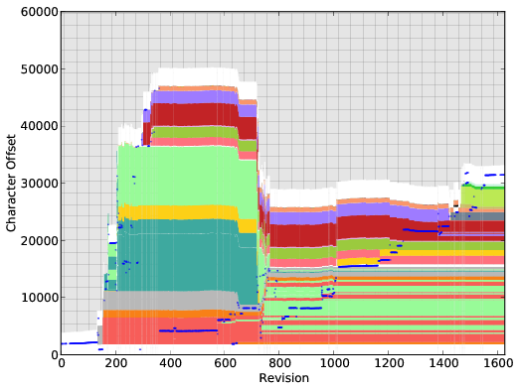
(b) Topic 54, Batch 2



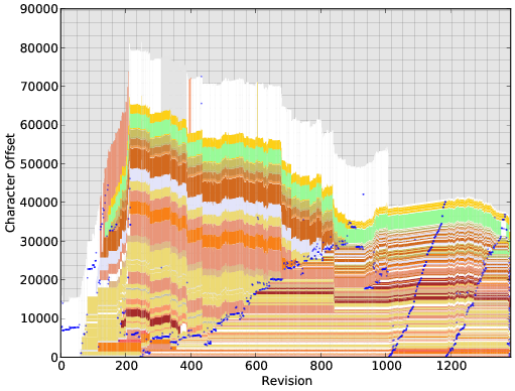
(c) Topic 103, Batch 2



(d) Topic 110, Batch 1



(e) Topic 79, Batch 2



(f) Topic 30, Batch 1

Figure 3.3: Visualizations of documents' editing histories.

exhibit primarily boil-down behavior. The remaining 77 (26%) can't be definitively assigned to either category—most of these documents exhibit both types of editing behavior in equal measure.

Aside from changes in the document's size and composition over time, the visualization also shows each revision's current edit location as a blue dot. This presents another interesting view into the data, in that different editing styles become apparent. Some authors write in a mostly linear fashion, either only ever adding on to the end of the document (Figure 3.3c), or giving the pasted source passages a single rewriting pass from front to back (Figure Fig. 3.3e). Others insert or edit passages in arbitrary places across the document (Figures 3.3a and 3.3b). While most authors edit each section of the document at most once, some revise the entire text once (Figure 3.3d) or even twice (Figure 3.3f) after writing a first draft.

In order to gain more insight into how strongly the type of editing behavior correlates with who wrote the corresponding document—and may thus embody individual working habits—we next look for patterns in the editing histories of each author's documents.

3.5 Aggregated Editing Behavior by Author

In order to get a more high-level view of the editing behavior, we consider only the top edges of the previous topic history plots—that is, the length of the document at a given point in time. We then normalize the axes, so that the x-axis shows the time as a percentage of the document's total revision history, and the y-axis shows the document's length as a percentage of the maximum length ever achieved throughout the editing process. This way, the editing histories for different documents can be easily compared, regardless of how much text they contain, and how often they have been revised.

Using these normalized document histories, we aggregate the work done by each author into a single figure by computing the mean of the resulting curves over all documents a given author has written. The result is shown in Figure 3.4. Here, the black line shows the author's average editing behavior, superimposed on the individual topics' normalized editing histories shown in grey. The x-axes show the relative revision number as a percentage of the total number of revisions for that topic; the y-axes show the relative character offset as a ratio of the maximum length the document attained throughout its history. The discontinuities in the curves for some of the topics result from software bugs during the corpus creation phase—in some cases, the editor the corpus authors used to write their texts deleted the entire document, which then had to be restored in a subsequent edit.

This aggregated view of authors' editing behavior allows for some interesting conclusions. Authors who favor the build-up style of text reuse tend to stick rather

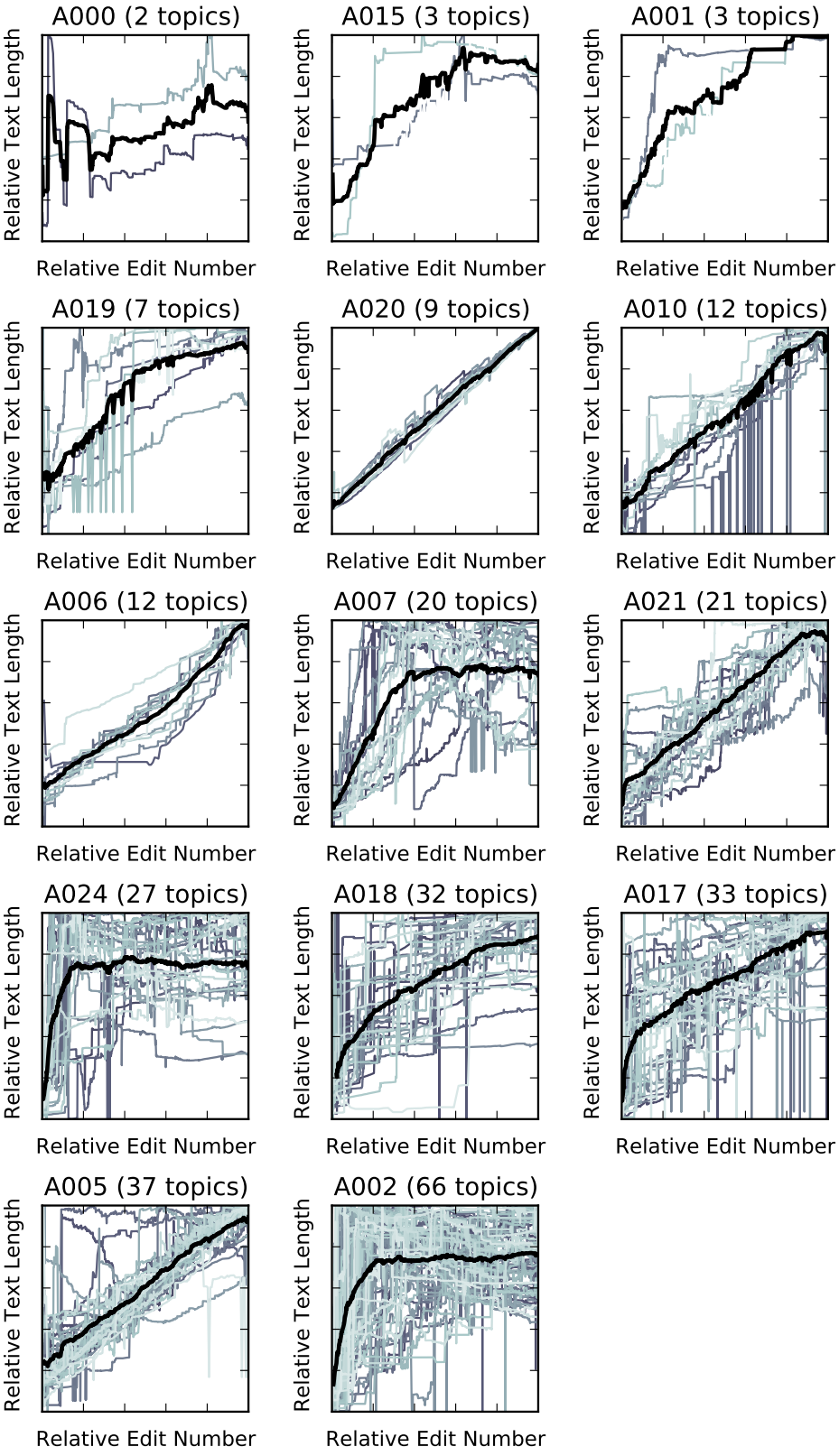


Figure 3.4: Aggregated editing histories for authors with 2 or more topics. Both axes show the interval $[0, 1]$. Individual topics are shown in tones of grey. The average of all topics is overlaid as a thick black line.

closely to this type of editing behavior. In the texts written by authors A020 and A006 (and to a lesser extent A010 and A021), document length tends to increase more or less linearly over time, and this is reflected in a linearly increasing average as well. For authors who prefer to reuse text in a more boil-down fashion—such as A002, A007 and A024—the average editing behavior tends to follow a pattern of logarithmic growth. Here, the average document length over time has a sharp increase at the beginning, followed by a long plateau.

The picture painted by individual documents' curves is much less clear for the boil-downers, since the point in the editing history where long text passages are inserted tends to vary widely. Some authors tend to alternate between both editing styles. As outlined in Section 3.4, both behaviors may even alternate within the same document. Author A005 favors build-up text reuse with only a few exceptions, resulting in a linear average editing behavior. The editing behavior of author A018 is spread over a much wider spectrum, appearing as a half parabola when averaged.

3.6 Summary

In this chapter, we examined how the documents in the corpus may be compared to each other and grouped into categories. Starting from a possible spectrum of corpus documents distilled from established taxonomies of plagiarism, we analyzed the degrees of paraphrasing in document passages and of interleaving material from different sources. We found the author dimension especially interesting in this regard, and detected hints of different working styles reflected in the documents' properties. An analysis of the way different authors modify texts over time provided even deeper insights into individual working habits.

While the subjects studied in this chapter provide interesting insights into the properties of the corpus documents and how they may yield different document categories, there is much room for future investigation. For instance, it seems promising to focus on correlating the edit history-based document properties explored in Sections 3.4 and 3.5 with the passage-level properties examined in the first half of this chapter. Even more data sources, such as the search engine logs examined in Chapter 4 may be called on to gain even deeper insights about the types of plagiarism that the corpus contains.

4 Search Missions for Source Retrieval

The Webis-TRC-12 dataset provides an unprecedented opportunity to study the relationship between user tasks and the resulting search behavior as observed in the query log, since the information needs behind the queries it contains are predetermined and well-known. As outlined in Section 2.4, each document written about one of the 150 topics in Batch 2 of the corpus (and its associated query log) is motivated by exactly one exploratory search mission. Beyond that, the corpus includes detailed information about the actual writing task in progress.

In this chapter, we compare the search missions in the Webis-TRC-12 query logs to exploratory search missions in a prior dataset. In order for our corpus to be useful in future search mission research, the properties of the query log and how they relate to previous efforts need to be studied. In Section 4.1 we compare our query logs to the Webis-SMC-12 dataset introduced in Section 2.5 at the surface level. We then formulate two hypotheses that relate to the novelty of the Webis-TRC-12 search missions and to the occurrence of exploratory search in the reference dataset.

In Section 4.2 we scrutinize a sample of the Webis-SMC-12 corpus for exploratory search missions. Our results indicate that while similar search missions do exist, our assumptions regarding the novel quality of our dataset seem to be justified. To study the relationship between the exploratory search missions in both datasets in more depth, we then devise an experiment to measure search mission similarity. Section 4.3 describes our set-up and results.

In the second half of the chapter, we tie into our analysis of writing behaviors presented in Chapter 3. Section 4.4 studies the time distribution of queries, and explores the possibility that authors' preferred editing strategies may be reflected in their interactions with the search engine. In Section 4.5 we study the relationship between querying and editing behavior at the level of individual topics.

4.1 Properties of the Webis-TRC-12 Search Missions

The search engine interactions recorded as part of Webis-TRC-12 present a new opportunity to study exploratory search behavior. In order to better understand their properties, we use Webis-SMC-12 as a reference dataset for comparison. We expect the search missions in the Webis-SMC-12 dataset to be closer to the distribution of typical search engine usage patterns in the real world, since it is constructed from a

representative sample of user interactions with a public search engine. This opens up the possibility to investigate how the Webis-TRC-12 search missions relate to normal day-to-day search activity.

As shown in Table 4.1, the properties of both corpora are quite different. Webis-SMC-12 contains many more search missions in total, but they are much shorter than the ones in our dataset. On average, a search mission in the Webis-TRC-12 dataset contains 25 times as many queries and 17 times as many clicks as a Webis-SMC-12 mission. While there are 370 search missions without any clicks in the Webis-SMC-12, no mission in our dataset has less than 10 clicks total. Webis-TRC-12 has almost three times as many queries in total. However, for 65.9% of the queries, none of the results was clicked on—a higher percentage than for Webis-SMC-12 at 51.7%.

In order to make the most future use of the Webis-TRC-12 query logs, it seems expedient to study how the patterns of user interaction recorded in them compare to the day-to-day use of a public search engine. It seems intuitively likely that in general, the spectrum of user interaction is much wider than the narrow range of search behavior of our authors seeking sources for text reuse. Our cursory analysis of the two datasets shown in Table 4.1 yields some evidence in favor of this intuition.

By contrast, we do expect to find some subset of real-world search behavior that exhibits properties similar to what our authors did. In fact, this is an important aspect of the usefulness of our dataset as a whole—if the types of interactions it contains were never observed in the real world, it would make little sense to study them.

We formulate these assumptions in the first of the following hypothesis:

Hypothesis 4.1. *The search missions in the Webis-TRC-12 query log are* ordinary *in the sense that the exploratory search behavior our authors exhibit can also be found in the Webis-SMC-12 dataset. At the same time, this type of search mission is* rare*, in that only a small subset of the wider spectrum of search missions in Webis-SMC-12 is similar to the ones found in Webis-TRC-12.*

As we expect the Webis-SMC-12 data to contain some subset of exploratory search missions, we are interested in the possibility of distinguishing exploratory and non-exploratory search missions based on their properties. We investigate the possibility of exploratory search missions in Webis-SMC-12 being similar to the Webis-TRC-12 ones:

Hypothesis 4.2. *Exploratory search missions in Webis-SMC-12 should be more similar to the Webis-TRC-12 missions, than to non-exploratory Webis-SMC-12 missions.*

In order to investigate the first of these two hypotheses, we apply a process of manual annotation of search missions.

Table 4.1: Statistics for the two query log datasets.

(a) Webis-TRC-12

Characteristic	**Distribution**					**Total**
	min	median	max	mean	σ	
Search Missions						**150**
Queries						**13 651**
- per Mission	4	64	616	91.01	83.41	
Unique Query Strings						**3 862**
- per Mission	1	22	122	25.89	18.75	
Clicks						**10 525**
- per Mission	10	56.5	305	70.17	50.84	
- per Query	0	0	76	0.77	2.20	
Unique URLs clicked						**7 273**
- per Mission	7	43	183	48.62	29.48	
- per Query	0	0	62	0.65	1.65	

(b) Webis-SMC-12

Characteristic	**Distribution**					**Total**
	min	median	max	mean	σ	
Search Missions						**1 393**
Queries						**5 091**
- per Mission	1	2	149	3.65	7.16	
Unique Query Strings						**3 715**
- per Mission	1	1	137	2.76	5.52	
Clicks						**5 885**
- per Mission	0	1	204	4.22	12.55	
- per Query	0	0	55	1.16	2.57	
Unique URLs clicked						**3 307**
- per Mission	0	1	94	2.79	6.39	
- per Query	0	0	34	0.89	1.67	

4.2 Exploratory Search Missions in the Reference Corpus

As outlined above, we expect only a small subset of the missions in Webis-SMC-12 to have characteristics of exploratory search missions; many of the information needs encountered in Section 2.4 can probably be met with a single query. The statistics of both query logs do indicate that this is the case. If we only consider the subset of the 20 longest search missions in Webis-SMC-12 by number of total interactions (queries and clicks), they have a mean query count of 40 and a mean click count of 90. This is much closer to the median query (click) count of 64 (56.5) in the Webis-TRC-12 dataset, and thus more similar to a typical search mission from the latter corpus. The fact remains, however, that Webis-TRC-12 missions tend to have fewer clicks per query than Webis-SMC-12 missions.

In order to test Hypothesis 4.1, we manually inspect the 200 longest search missions in Webis-SMC-12. Out of these, we can only positively identify 10 that show definite properties of exploratory search. Table 4.2 shows an example of a mission in Webis-SMC-12 that we identified as exploratory. The second half of the table shows one of the Webis-TRC-12 missions for comparison. While different in subject matter, both missions show similar properties in several respects. Queries tend to explore a subject from multiple angles and are reformulated frequently. Multiple result documents are clicked on.

Both missions span several weeks, and can be divided into a number of sessions with longer breaks in between. While the time cutoff to delineate sessions is a matter of debate, we choose the value of 90 minutes to split physical sessions, as advocated by Hagen et al. (2013). The session breaks are highlighted as dashed lines in Table 4.2.

Having identified 10 out of 1 393 search missions as showing exploratory characteristics, we have shown both the existence and rarity of exploratory search missions in Webis-SMC-12. This can be counted as evidence in favor of Hypothesis 4.1. However, in order to answer Hypothesis 4.2, we must quantify the similarity of exploratory search missions in both datasets. To this end, we propose a set of search mission features, and a simple experiment to measure the similarity of search mission datasets.

4.3 Measuring the Similarity of Search Missions

In order to be able to quantify search mission similarity, we first define a simple set of features to describe the properties of a search mission. The features we choose are partially inspired by previous attempts at modeling the properties of search missions, such as the ones by Agichtein et al. (2012) and Kotov et al. (2011), though our use case is very different.

Table 4.2: Two similar search missions from the Webis-TRC-12 and the Webis-SMC-12 corpora. The first column in each table shows the time elapsed since the previous query. Dashed lines denote likely session breaks.

(a) Webis-SMC-12

Time gap	Query	Clicks
(n/a)	quartiles	4
4 days, 1:06:19	sohcahtoa	0
57 days, 0:58:45	probability	1
0:01:20	normal probability	4
1 day, 20:25:41	probability	0
0:00:28	what is the probability	1
0:03:14	using the normal curve to find probability	1
0:19:21	what is the probability that the value will fall within	3
0:17:58	what is the probability	0
0:19:02	what is the probability	0
0:04:51	what is the probability..will fall within	1
0:02:22	what is the probability..will fall within	0
1:59:13	hypothesis testing	1
0:00:49	two tail hypothesis testing	1
0:35:35	sample size	6
0:17:13	calculate the sample size	8
10:15:11	finding the p-value	2
0:03:02	proportion for 2populations	0
0:00:45	proportion for 2 populations	1
0:11:25	linear regression	1
0:05:39	find the p value	1

(b) Webis-TRC-12: Topic 112, Batch 2

Time gap	Query	Clicks
(n/a)	Gas Water Heater	1
0:00:10	Gas Water Heater	2
0:03:39	Gas Water Heater	1
0:00:51	Gas Water Heaters	1
0:00:10	Gas Water Heaters	3
19 days, 20:31:46	water heaters	1
0:08:06	gas heating	3
0:09:27	thermal water heating	1
0:59:11	electric water heaters	0
0:03:20	solar water heaters	1
0:00:31	solar water heaters	1
0:01:44	heating water	1
0:02:07	gas heating	3
0:04:58	thermal water heating	1
0:00:21	thermal water heating	0
0:00:26	thermal water heating	1
1 day, 20:52:47	Kenmore Gas Water Heaters	0
0:00:03	Kenmore Gas Water Heaters	3
0:08:09	gas water heaters	2
0:27:16	cost comparison gas to electric water heaters	1
0:00:23	cost comparison gas to electric water heaters	0
2 days, 23:03:38	geysers	2
0:00:04	geysers	0

Table 4.3: Mission-based, query-based and pair-based features used in our experiment.

Mission-based	**Query-based**	**Pair-based**
Distinct query strings	Click count	Activity gap
Distinct URLs	Highest rank clicked	Edit distance
Session count	Characters	

Table 4.3 summarizes the features we implement. The mission based features are computed once per search mission, and include the number of distinct query strings submitted and URLs clicked throughout the search mission. In addition, we compute the number of sessions, using the 90-minute heuristic noted above.

Query-based and pair-based features are computed for each query in the search mission, and for each pair of consecutive queries, respectively. The former include the number of clicks on a query's search result list, the result rank of the result clicked, and the number of characters in the query string. For query pairs, we compute the time in between their occurrence, and the Levenshtein edit distance between the respective query strings. Since a search mission potentially contains many queries and query pairs, we take the medians of these values to characterize the search mission.

In order to use these features in our search mission similarity experiment, we derive four datasets from the Webis-TRC-12 and Webis-SMC-12 query logs. In addition to the 10 exploratory search missions identified in the Webis-SMC-12 query logs, we found another 16 that couldn't be definitively decided either way. For the purpose of our experiment, we consider these borderline cases separately. Beyond that, not all of our features can be computed for every search mission in the Webis-SMC-12. For instance, the pair-based features can only be computed when there are at least two queries. We hence discard 832 search missions that are too short to compute all of the features.

Of the remaining 561 search missions from Webis-SMC-12, 10 are considered to be exploratory search missions. For the following experiment, these exploratory search missions form the dataset SMC_e, and the 535 non-exploratory missions, the dataset SMC_n. The 16 borderline-exploratory search missions form the dataset SMC_b. We will refer to the 150 exploratory search missions in the Webis-TRC-12 query logs as dataset PC.

To measure the similarity between two search missions, we represent each as a vector of its feature values, and compute the Euclidean distance between the two vectors. Pairs of search missions with a low Euclidean distance are considered similar. Since the numerical magnitudes of the features are quite different, we apply mean normalization before computing the Euclidean distance. This normalization strategy involves computing the mean and standard deviation of each feature on the union of

Table 4.4: Pairwise similarity of exploratory and non-exploratory search missions in our four datasets.

Datasets		**Mission pairs**	**Pairwise Euclidean distance**				
			min	median	max	mean	σ
PC	SMC_e	1 500	0.48	3.43	17.03	4.07	2.32
PC	SMC_b	2 400	0.83	3.84	16.87	4.27	2.17
PC	SMC_n	80 250	0.65	4.19	27.08	4.79	2.46
SMC_e	SMC_n	5 350	0.65	2.85	23.89	3.49	2.28
SMC_e	SMC_b	160	0.92	2.86	9.28	3.25	1.85
SMC_b	SMC_n	8 560	0.40	3.17	24.03	3.28	1.59

all four datasets, and then subtracting the vector of means from each feature vector and dividing by the vector of standard deviations. This ensures that all features have similar magnitudes—so that no single feature will dominate a vector and hence the Euclidean distance—but preserves the shapes of individual features' distributions.

We examine the similarity of the four datasets we derived above by computing the pairwise Euclidean distance of all pairs of search missions for each pair among the four datasets. Table 4.4 shows the distribution of pairwise distances.

As is apparent, our results do not fully confirm Hypothesis 4.2. Per the hypothesis, we would have expected a higher similarity between the PC and SMC_e datasets than between SMC_e and SMC_n. However, both the median and mean distances are lowest for the SMC_e-SMC_n-pair. Nevertheless, the distance between search missions in PC and SMC_e tends to be much lower than between those in PC and SMC_n. For the borderline dataset, the average similarity to PC search missions does indeed fall in between those for SMC_e and SMC_n. However, the most similar search mission pairs in PC and SMC_b are actually less similar than those in PC and SMC_n.

Taken as a whole, these results support a somewhat weaker version of Hypothesis 4.2: that Webis-TRC-12 search missions are much more similar to exploratory search missions, than to non-exploratory search missions from the Webis-SMC-12.

Aside from comparing the search missions in the Webis-TRC-12 query log to other search mission datasets, we can also join together the different data sources within Webis-TRC-12. The remainder of this chapter studies the time distribution of search engine queries, and its relationship to authors' writing activity.

4.4 Percentage of Queries Submitted Over Time

In Sections 3.4 and 3.5 we explored the time dimension of document editing histories. The change in document length over time proved especially interesting in this regard, and showed some correlation with individual authors' working habits.

A similar analysis is possible for the query logs in the Webis-TRC-12 dataset: we can measure the number of queries that have been submitted up to a given time as a percentage of the total number of queries for that topic. If we view the time dimension as a fraction of the total time the author spent working on that topic, the result is a function with a domain and range of $[0, 1]$.

Figure 4.1 shows plots of the queries-over-time function for all 150 topics in Batch 2 of the corpus. In each individual graph, the y-axis shows the cumulative percentage of queries submitted up to the time shown on the x-axis. The time itself is measured as a percentage of the total time spent working on a given topic. To make comparisons between different plots more meaningful, we normalize breaks in the query log before plotting. All periods of inactivity are clipped to 5 minutes, i.e., breaks of more than 5 minutes are limited to that threshold.

The plots for individual topics in Figure 4.1 are ordered by decreasing area under the curve and show considerable differences in the distribution of queries over time. For instance, topic 047 in cell A1 has 90% of its queries submitted during the first 20% of working time. For topic 100 in cell J15, the author has submitted less than 10% of all 64 queries after 60% of the time has passed.

Topics in the middle rows of Figure 4.1 fall between these two extremes. Some of them, such as topic 133 in cell A10 and topic 002 in cell F11, show an approximately linear relationship between time spent and queries submitted. In other cases, queries occur in several bursts throughout the time spent working on the topic, which results in a number of steps in the curve. This is true for topic 146 in cell F9, and topic 060 in cell I10, for example. Note that all of these topics have 60 or more queries—the steps are not caused by sparse data. Still other types of behavior exist: for topic 050 in cell F12, close to 30% of the 78 queries are submitted in a burst right at the beginning, followed by a more linear querying behavior after 60% of the time has elapsed.

The time distribution of queries for individual topics gives a good overview of the wealth of different search strategies found in the Webis-TRC-12 query log. Next, we study the aggregated querying behavior over several topics. This may provide new insights into typical search strategies employed by different authors.

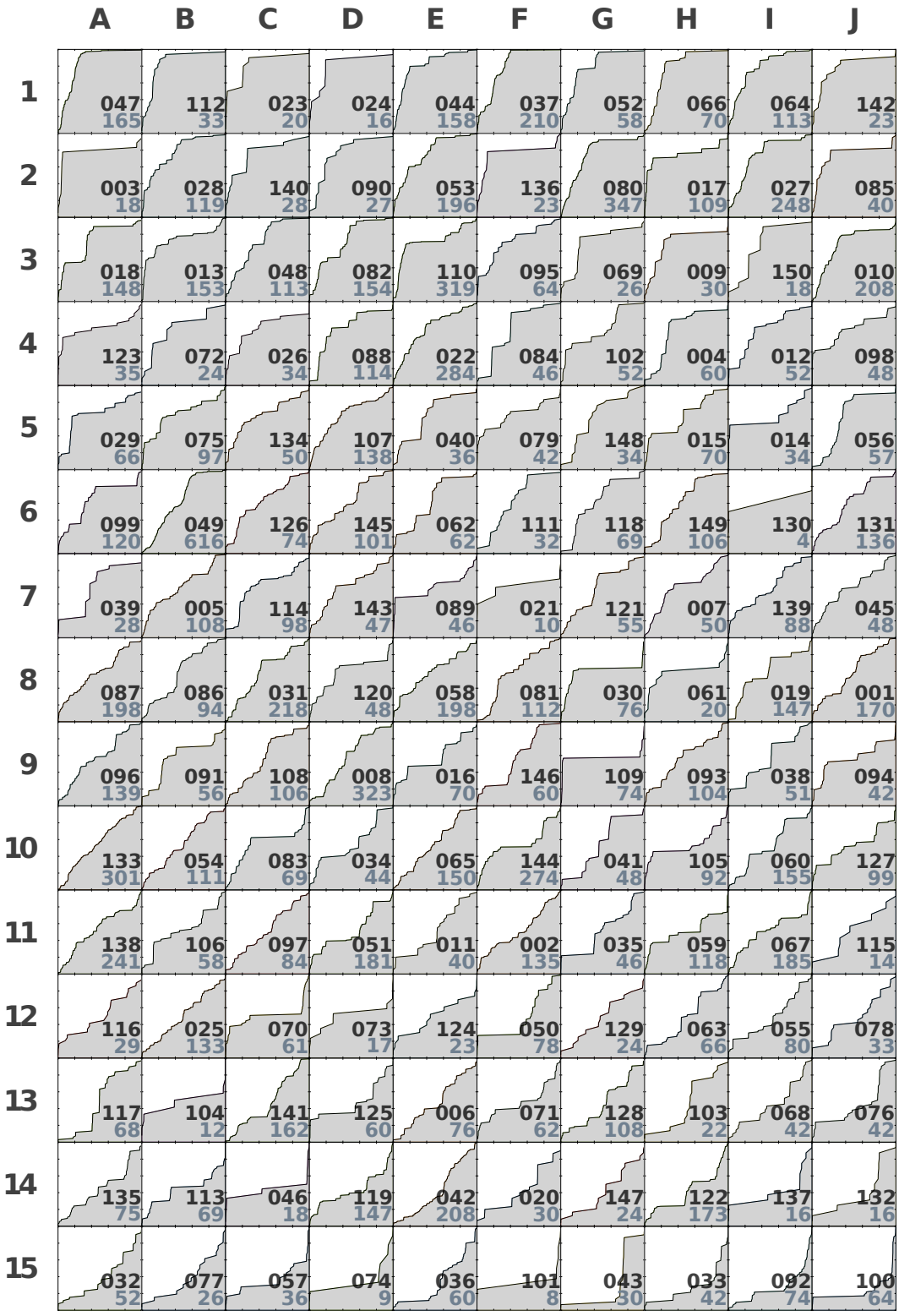


Figure 4.1: Percentage of queries submitted over time for all 150 topics in Batch 2 of Webis-TRC-12. The filled grey curve shows the cumulative number of queries submitted as a function of time spent working on a topic. The topic number is overlaid in black, the number of queries for each topic in dark grey. Plots are ordered by area under the curve.

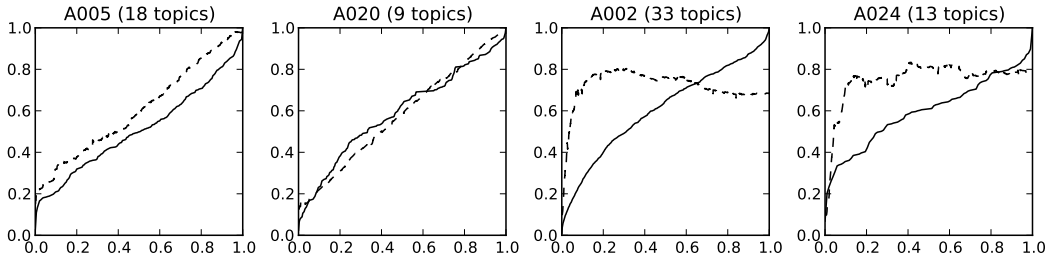


Figure 4.2: Correlation of average editing and querying behavior for a selection of four authors. The solid line represents the percentage of queries submitted over time aggregated over all the author's topics shown in Figure 4.1. The dashed line corresponds to the average editing behavior shown in Figure 3.4.

4.5 Correlating Querying and Editing Behavior

In a similar vein to the aggregated edit history plots shown in Section 3.5, we can aggregate the querying behavior over several topics. This allows insights into the average time distribution of queries for the topics a given author has written. It also allows us to study the correspondence of querying and editing behavior. In Section 3.4 we observed two main editing strategies: boil-down, where most of the content is pasted in at the beginning, and then rewritten, and build-up, which follows a more linear strategy of pasting and editing consecutive passages.

In Figure 4.1, we find a similar situation with respect to the time distribution of queries—in some cases, most of the queries were submitted near the beginning of the time span during which the author worked, in others, queries were submitted late, or are distributed evenly over time. We expect a connection with the editing behavior, and formulate the following hypothesis:

Hypothesis 4.3. *There is a correlation between querying and editing behavior. Authors who favor a boil-down style of text reuse submit many queries early on; the queries for authors who tend towards a build-up approach will be distributed more evenly over time.*

Confirming this hypothesis may imply that it is possible to predict editing behavior from the characteristics of the query log and vice-versa.

As a simple test, we visually analyze the distributions of edits and queries. Figure 4.2 shows a comparison of the average editing and querying behavior for a sample of four authors. The solid line shows the average query distribution for all the author's topics; this is simply the mean of the corresponding curves from Figure 4.1. The dashed line shows the mean document length, similar to Figure 3.4; however, only the Batch 2 topics were averaged in the present case.

The results of this initial investigation are inconclusive. Authors who strictly follow a build-up editing style tend to show linearly increasing query distribution, as seen in the first two plots in Figure 4.2. However, for authors who favor the boil-down approach, or whose editing style tends to vary from topic to topic, there is much discrepancy. The latter two plots in Figure 4.2 show examples of this. Out of all 10 authors who wrote at least two documents for Batch 2, five show a clear correlation between editing and querying, and the other five do not.

As we found in Section 3.5, build-up authors tend to firmly stick to their editing style, whereas those who sometimes work in a boil-down fashion tend to be more flexible. This fact may explain our ambiguous results with respect to Hypothesis 4.3, but it does not rule out the possibility that a correlation exists. We can further examine our hypothesis if we drill down to the topic level.

Rank Correlation at the Topic Level

As already shown in Figure 4.1, the time distribution of queries implies a total order of the list of topics by the area under the query distribution curve. Topics with most queries submitted at the beginning have a larger integral than those where queries occur later. Arguably, the same is true for the normalized editing histories as shown in Figure 3.4. If we order the topics by the area under the corresponding length-by-revision curve, texts that receive most content early on will be ranked higher than those that are extended in a more linear fashion.

These considerations yield two different rankings of all the topics in Batch 2. A high similarity of these orderings would be evidence in favor of Hypothesis 4.3. To measure the similarity of these ranked data, we employ several correlation coefficients. As noted by Everitt and Skrondal (2010, p. 107), a correlation coefficient is a statistical index to quantify the linear relationship between two variables. Multiple different formulations of correlation coefficients exist, but they all share some basic properties: the value of the correlation coefficient is in the range $[-1, 1]$, where the sign indicates the direction, and the magnitude the strength of the linear relationship; a value of zero implies no correlation between the tested variables.

In order to investigate the relationship between the query-based and the edit-based rankings, we apply the Kendall's τ, Spearman's ρ, and Pearson's product moment correlation coefficients. In each case, we use the implementation in the R statistical programming environment.[1] We observe a Kendall correlation of 0.132, a Spearman correlation of 0.199 and a Pearson correlation of 0.197. While this result does not force us to reject Hypothesis 4.3 altogether, as Zou et al. (2003) explain, it is at best weak evidence of a positive correlation between the two measures.

Given the measures we explored in this chapter and our current understanding of the data, we cannot conclusively prove a correlation between editing and search

[1] `http://www.r-project.org/` (last accessed March 2013)

strategies for every document in the corpus. However, while the simple measures we devised do not yield a clear answer, a more in-depth analysis may do so. For instance, a semantic analysis of query terms, the content of source documents clicked, and the text changed in a given edit may reveal a causal relationship between events in the query log and edit log. This may suggest a different ordering of queries versus edits. The temporal ordering explored in this chapter may not always be accurate—such as in cases where an author gathers material for multiple separate aspects of the topic in parallel.

4.6 Summary

In this chapter, we studied the nature of the exploratory search missions in Webis-TRC-12. Using the Webis-SMC-12 search mission dataset as a reference, we gained some insights into the prevalence and properties of exploratory search in a dataset curated from a public search engine. We devised a simple method to compute the similarity between search missions, and found some evidence that exploratory search missions can be quantitatively distinguished from non-exploratory ones. Our work is only a first step—better similarity measures, and more in-depth studies may yield a much clearer picture.

In the latter part of the chapter, we made a connection to the study of revision histories shown in Chapter 3. While for many topics the nature of the relationship with the query log remains elusive, we found a strong correlation between build-up editing and a linear search strategy. At the very least, our findings demonstrate the suitability of Webis-TRC-12 for studying search behavior.

White et al. (2008) summarize a range of studies evaluating exploratory search systems. Their survey shows that there is currently no standard corpus of exploratory search data to support such an evaluation. Webis-TRC-12, with its comprehensive query logs supporting long-term writing tasks, may help fill this gap in the future.

5 Corpus Construction and Statistics

The previous two chapters have delved into selected aspects of Webis-TRC-12 in considerable depth. By contrast, this chapter showcases additional characteristics of the data from a more distant view. We give insights into the corpus creation process and the statistical properties of the data, and answer some basic questions that arise. In the process, we highlight some gaps in our understanding of the data that may be of interest for future research.

In Section 5.1, we examine the crowdsourcing effort that produced the documents in Webis-TRC-12. We discuss the distribution of the work among the different authors, author demographics and the financial cost of the crowdsourcing effort. Section 5.2 explores some additional dimensions of the corpus documents, such as the relationship between document length and the way authors retrieved their sources.

Section 5.3 showcases additional insights gained from the query log. Most notably, it highlights how information about which documents authors clicked on and which documents they used as sources can be interpreted as relevance judgements. Finally, Section 5.4 introduces some interactive software tools that we developed to further our (and others') understanding of the data.

5.1 Corpus Generation Through Crowdsourcing

The starting point for any corpus construction effort is deciding how corpus documents should be created. As outlined in Section 2.1, there are three fundamental sources of texts for constructing a corpus of plagiarized text: artificial, simulated, or real plagiarism. Since Potthast et al. (2012a) chose the middle-road strategy of simulated plagiarism, they found themselves in need of a large number of human annotators to write corpus documents.

In order to attain a high degree of quality and realism in the corpus documents, all writers needed to be fluent in English, and experienced in writing English text. Due to the considerable time investment needed to produce a large number of documents of the desired length and quality, crowdsourcing was chosen as the corpus construction approach. While some of the previous corpora mentioned in Section 2.3 also made use of crowdsourcing, Webis-TRC-12 breaks new ground in terms of task complexity: writers had to work on an entire document of several thousand words, rather than just short passages.

Table 5.1: Authors and topic assignments for Webis-TRC-12.

	Batch 1	Batch 2	Overall
Topics	147	150	297
- volunteer	13	0	13
- paid	134	150	284
Authors	24	12	27
- volunteer	10	0	10
- paid	14	12	17
Topics per author			
- minimum	1	1	1
- median	1.5	11.5	2
- maximum	33	33	66

The remainder of this section examines various aspects of the crowdsourcing effort, starting from the distribution of the work among individual authors.

Authors and Topic Assignments

As pointed out in Section 2.2, Potthast et al. (2012a) divided the corpus creation effort into two distinct batches of 147 and 150 documents. The main difference between both is in the way authors retrieved source documents for plagiarism. While Batch 1 authors chose from a list of ClueWeb09 documents judged as relevant to the topic in previous TREC competitions, in Batch 2 they retrieved their own sources using the ChatNoir search engine. This explains the difference in number of topics per batch—for three out of the 150 TREC topics used in the corpus creation effort, none of the relevant sources were available.

Table 5.1 gives an overview of the worker-topic-assignment for both batches. As mentioned in Section 2.2, the main purpose of the batch separation was to control for the effects of different retrieval models on source selection. However, Batch 1 also served as a testbed for the text writing interface. To ensure all of the technology involved in the corpus creation was running smoothly, 13 of the documents in Batch 1 were written by 10 volunteers recruited from university staff. Paid authors recruited from the crowdsourcing platform oDesk wrote the vast majority of the documents in both batches, however. In total, 27 different authors wrote texts for the corpus, 17 of them professional writers.

Additional detail is shown in Figure 5.1, which shows the exact number of topics per batch that each author was assigned. It is apparent that the number of documents written by each author varies considerably—the most prolific of the workers wrote more than 20% of all documents, while many others only wrote a single one. Due

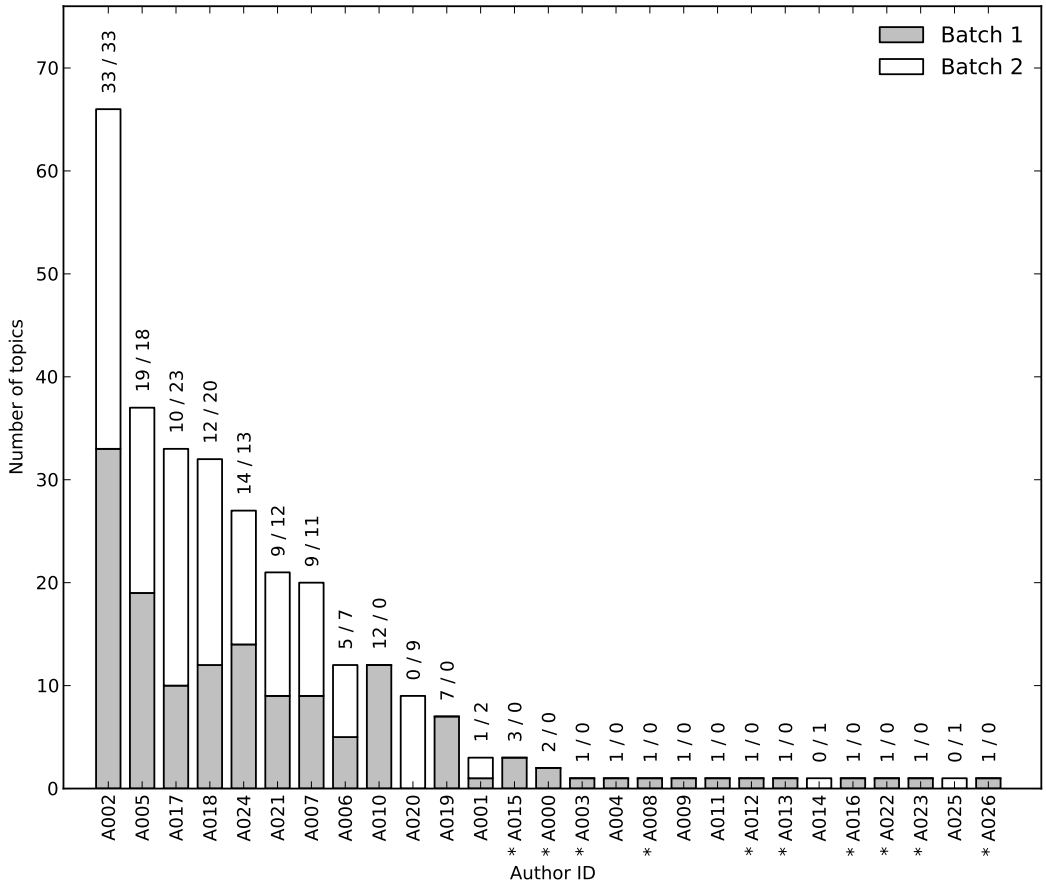


Figure 5.1: Number of topics per author for the two batches. Volunteer authors are marked with an asterisk.

to the lack of volunteer writers, topics are more evenly distributed in Batch 2 over fewer authors overall. Most of the paid authors wrote documents for both batches, with the exception of four who participated only in the first batch, and three who participated only in the second batch. While topics repeat across batches, no author was assigned the same topic twice.

The five most prolific authors wrote two thirds of all documents in the corpus, while the top ten authors divide over 90% of the documents amongst themselves. This fact is interesting for investigations of author behavior: it may be expedient to focus such studies on these most productive authors—as we have done in Chapters 3 and 4 of this thesis.

Table 5.2: Demographic survey of the 12 authors in Batch 2.

Author Demographics			
Age		*Gender*	
Minimum	22	Female	67%
Median	37	Male	33%
Maximum	65		
Country of origin		*Country of residence*	
Australia	8%	Australia	8%
India	17%	India	17%
Philippines	25%	Philippines	25%
South Africa	8%	South Africa	17%
		Sweden	8%
UK	25%	UK	8%
USA	17%	USA	17%
Professional writing experience			
Years		*Genre(s)*	
Minimum	2	Web content	25%
Median	5	Business	25%
Maximum	20	Medical	17%
Mean	8.2	Fiction	17%
Standard dev.	5.85	Other	25%

Academic degree		*Native language(s)*	
Postgraduate	33%	English	67%
Undergraduate	25%	Filipino	17%
None	17%	Hindi	17%
n/a	25%	n/a	8%
Second language(s)		*Search engines used*	
Afrikaans	8%	Google	92%
Dutch	8%	Bing	33%
English	33%	Yahoo	25%
French	17%	Others	8%
German	8%		
Spanish	8%	*Usage*	
Swedish	8%	Daily	83%
None	8%	Weekly	8%
n/a	8%	n/a	8%
Reuses text		*Has plagiarized before*	
Often	17%	Yes	33%
Sometimes	42%	No	59%
Never	33%		
n/a	8%	n/a	8%

Author Demographics

The original job posting on the oDesk platform sought professional writers with experience writing in English. While all of the hired authors speak English fluently, they come from a variety of geographic and language backgrounds. In Batch 2, the writing task included an optional questionnaire that asked a number of demographic questions. The results are summarized in Table 5.2.

As shown in the table, the typical author is in their thirties and well-educated. Two thirds of the authors are female. The most frequent countries of origin for our authors are the Philippines and the UK, followed by the US and India. All authors have two or more years of professional writing experience, predominantly in non-fiction genres.

Two thirds of the authors speak English as a first, the remaining third as a second language. Filipino and Hindi are named as first language by two authors each; two authors reported more than one first language. Among second languages other than English, French is most common. Only one author reported not speaking a second language.

The remaining questions focused on authors' experiences with search engines and reusing text. Almost all authors report using Google most frequently; the vast majority avail themselves of web search engines daily. While almost two thirds of

the authors claim to reuse text, one third admit to having committed plagiarism in the past.

Author Feedback On Individual Topics

In addition to surveying authors for basic demographic information, Potthast et al. (2012a) included a list of topic-related questions with each of the 150 topics in Batch 2 of the corpus. Some of these questions were to be answered before starting a topic, and some of them after completing it. In the questionnaire, authors were asked to judge aspects of the topic such as the difficulty of finding relevant sources and of the topic itself. In each case, they were asked to compare their preconceptions before starting work on a topic, to their opinion afterwards. In another part of the survey, authors provided some basic information about their approach to plagiarizing text. This includes the number of paragraphs they rewrote and rearranged, the amount of content they reused from Wikipedia, as well as how easily they thought their text reuse could be detected—by a human or a machine, with or without access to the original sources.

Table 5.3 summarizes the results of this survey. Even in aggregated form, they provide some interesting insights into the authors' engagement with their task. For instance, the self-reported level of expertise on the subject of the plagiarized text tends to be much higher after writing it. This is perhaps somewhat surprising, since intuitively, plagiarizing a text seems to imply a lower level of involvement with the subject than writing one from scratch.

Also notable is the fact that authors seem to judge the capability of machines to detect plagiarism higher than that of humans. Only one author assigned a high likelihood to the possibility of a human without access to source documents detecting their plagiarism; 43 did so for a machine in the same situation.

The remaining questions are of interest to future studies like those presented in Chapters 3 and 4. Trying to correlate a topic's perceived difficulty with editing and querying behaviors may yield additional insights, as may the relationship between an authors' reported amount of rewriting/rearrangement and the paraphrasing-interleaving spectrum explored in Chapter 3.

Crowdsourcing Expenses

Crowdsourcing is far from the only possible approach to corpus creation, but—depending on the desired size, number, complexity and novelty of documents—it may turn out to be the only viable one. The investment of time and resources this entails then becomes a major concern during the planning phase of a new corpus. The Webis-TRC-12 dataset breaks new ground in terms of how long and complex

Table 5.3: Results of the author survey on individual topics. Each of the 150 topics in Batch 2 included a questionnaire.

About the topic					
Subject knowledge	**Expert**	**Much**	**Little**	**None**	**N/A**
- before	1	17	85	40	7
- after	8	103	20	3	16
Topic difficulty	**Experts only**	**Medium**	**Laymen**		**N/A**
- expected	21	86	36		7
- with hindsight	25	74	37		14
Difficulty finding sources	**High**	**Medium**	**Low**		**N/A**
- expected	16	93	31		10
- with hindsight	30	69	38		13
About the plagiarism					
Modified paragraphs	**None**	**Some**	**Most**	**All**	**N/A**
- rewritten	10	30	66	26	18
- rearranged	0	54	25	54	17
Wikipedia content	**None**	**0-10%**	**10-30%**	**30-90%**	**N/A**
- percent of total words	77	17	16	17	23
Expect detection	**Yes**	**Maybe**	**Prob. Not**	**No**	**N/A**
- by human given sources	63	42	27	4	14
- by human w/o sources	1	23	83	29	14
- by machine given sources	70	50	12	3	15
- by machine w/o sources	43	43	35	15	14

its documents are. Hence, it may serve as a kind of pilot study to inform similar efforts in the future.

The crowdsourcing platform oDesk, where all paid authors for the corpus were recruited, provides the employer with detailed statistics about hours worked and wages paid. Workers on oDesk are essentially freelancers who set their own hours and wages; hourly pay varies greatly depending on the individual's experience and country of residence.

Table 5.4a summarizes the hours and wages for 16 of the 17 authors recruited on the oDesk platform (one author of a single document is not included, due to the data curation effort still being ongoing at the time of this writing). As outlined above, Potthast et al. (2012a) chose a variety of authors from different backgrounds and levels of experience. Correspondingly, their hourly salary varies on a range between 3 and 34 US dollars, with a typical hourly rate being around 11 dollars. The choice of how many texts to write—as well as how much effort to put into rephrasing plagiarized passages—was left up to the individual author. While a few worked less than ten hours total, one author spent nearly 700 hours working on the corpus.

All told, completing the writing tasks for the 284 crowdsourced corpus documents required over 2 000 working hours and cost more than 20 000 US dollars.

Given the data provided by the crowd sourcing platform and the length of corpus documents, we can also examine the unit costs incurred. Table 5.4b shows the distribution of time and money invested per document, sentence, and word. In the Webis-TRC-12 corpus creation effort, the time required to complete a single corpus document varied between 3 and 13 hours, with a median document needing about 7 hours of work.

Due to the pioneering nature of the corpus creation effort, the return on investment is probably far from optimal. One major goal of the corpus creators was to study a wide variety of different authors and working styles. The range of observed data indicates that a future effort of similar nature—when focused less on diversity and more on minimizing cost—may be able to procure corpus documents for as little as 11 USD apiece.

This foray into the economics of corpus creation concludes our discussion of the crowdsourcing aspect of Webis-TRC-12. We next discuss additional dimensions found in the data that have not been the focus of attention thus far.

5.2 Document Statistics and Source Retrieval Models

As described by Potthast et al. (2012a), authors wrote corpus documents in a web-based rich-text editor. The editor stores a new revision of the document in a server-side Git repository every time the user stops typing for more than 300ms. The fine-grained insights this allows into the revision history of the plagiarized documents were amply explored in preceding chapters. This section enumerates some additional variables present in the documents and their revision histories, and studies their statistical properties.

Document Length

Table 5.5 shows the distribution of document lengths in words and sentences across the corpus. As is apparent from these data, the documents in Batch 2 tend to be somewhat longer. In order to provide a rough target, Potthast et al. (2012a) instructed authors to write around 5 000 words on each topic. Out of all 297 corpus documents, 88% hit at least 90% of that target word count; 94% of the documents have at least 2 500 words. Batch 1 authors—given a limited set of predetermined sources—seemed to have somewhat more difficulty meeting the word count requirement than those in Batch 2 who could find their own sources.

To examine the statistical strength of the influence of batch assignment on a document's length in words, we can employ a paired difference test. To that end, we

Table 5.4: Financial data for the Webis-TRC-12 crowdsourcing effort.

(a) Hours and salaries for oDesk workers.

Author	**Hours**	**Salary (USD)**	
		Per hour	**Total**
A001	15.67	11.11	174.06
A002	679.00	5.56	3 775.24
A004	13.17	11.11	146.28
A005	149.00	18.89	2 814.61
A006	112.00	5.56	622.72
A007	78.67	6.67	524.71
A009	8.33	10.00	83.33
A010	43.67	4.44	193.88
A011	3.33	3.33	11.10
A014	9.83	34.00	334.33
A017	309.17	14.43	4 460.48
A018	232.67	13.00	3 024.67
A019	39.17	7.78	304.72
A020	43.67	27.78	1 213.06
A021	200.83	6.67	1 339.56
A024	130.17	11.11	1 446.15
Median	61.17	10.55	573.71
Mean	129.27	11.96	1 279.31
Std. dev.	167.25	8.23	1 399.87
Sum	**2 068.33**		**20 468.90**

(b) Distribution of investments for the 284 documents written by paid workers.

	min	**median**	**max**	**mean**	σ
Hours / Document	3.333	7.271	13.167	7.309	2.651
Hours / Sentence	0.006	0.031	0.667	0.040	0.053
USD / Document	11.10	57.20	334.33	72.33	36.59
USD / Sentence	0.037	0.274	5.071	0.401	0.553
USD / Word	0.001	0.011	0.258	0.017	0.026

Table 5.5: Length distribution of the 297 corpus documents.

	(a) Overall		(b) Batch 1		(c) Batch 2	
	words	sentences	words	sentences	words	sentences
mean	5 594.4	242.7	5 315.2	234.3	5 868.0	251.0
std	1 576.0	80.3	1 722.4	90.5	1 369.3	68.2
min	260.0	14.0	260.0	14.0	1 205.0	42.0
median	5 803.0	247.0	5 765.0	242.0	5 828.5	252.0
max	15 569.0	714.0	9 504.0	482.0	15 569.0	714.0

view the set of of 147 topics that occur in both batches as the sample population. For a given topic, we view the length of the document in Batch 1 and the document in Batch 2 as two different measurements of the same sample.

In this scenario, Everitt and Skrondal (2010, p. 271) recommend a matched pairs t-test to evaluate the hypothesis that the means of both measurements differ. We employ the implementation of Student's two-sided t-test for paired samples provided in the SciPy library.[1] This test assigns a probability (the p-value) to the alternative hypothesis that the means of the populations of both measures are the same. Given the length distribution of documents summarized in Table 5.5, we arrive at a p-value of 2×10^{-3}. That is, the means of the word counts across batches are significantly different with a confidence of 99.8%.

While different explanations for this significant difference in document length are conceivable, we consider the fact that Batch 1 authors were limited in their source selection to be the most likely explanation. A more in-depth study of this effect is left for future work.

Figure 5.2 shows a box-and-whisker diagram of the distribution of document lengths by author. Medians are shown as a vertical black line and the interquartile range (centered around the mean) as a grey box. The whiskers extend to 1.5 times the interquartile range; data outside this range are shown as flier points.

As is apparent from the figure, the distribution of the document length varies somewhat for different authors. Four of the five longest documents in the corpus were all written by the same person (Author A007, some outliers not shown for scale). Only five of the authors wrote documents shorter than 2 000 words, and no author produced a set of documents with a median word count lower than 2 000. For the most prolific author, A002, document lengths are distributed almost across the entire range exhibited by the remaining authors. A number of documents are considerably shorter than the 5 000-word target. Most of these are in Batch 1, where source availability was the limiting factor.

[1] `http://www.scipy.org` (last accessed March 2013)

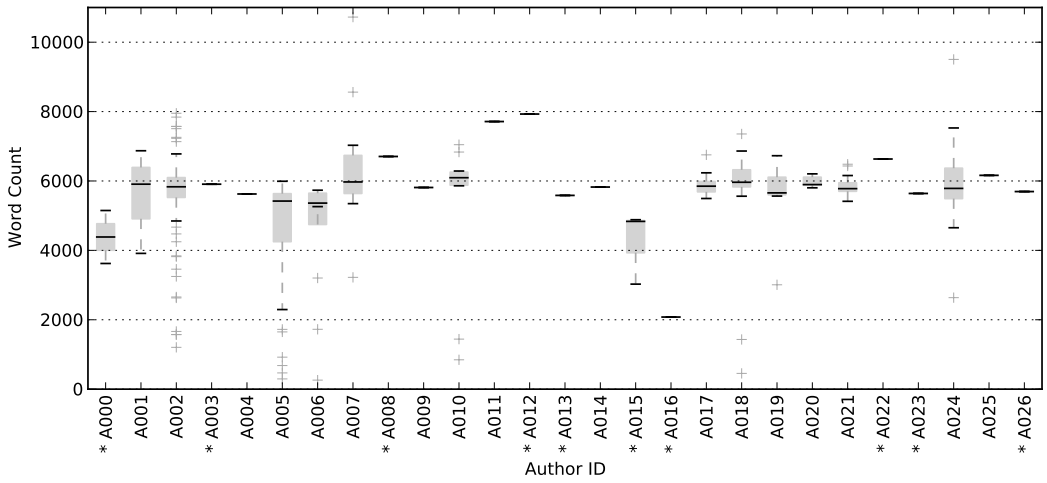


Figure 5.2: Distribution of document length in words by author. Volunteer authors are marked with an asterisk.

There is no immediately obvious connection between the author being hired or a volunteer, and the length of the document. While the document written by one of the volunteers is very short by other authors' standards, two others produced comparatively long texts. The remaining volunteers lie mostly in the middle of the length spectrum. A more in-depth study of the properties of volunteer versus paid documents may be of interest for future work.

Number of Revisions per Document

To measure the amount of work that went into producing a document, as well as the amount of obfuscation we can expect for a document's plagiarized passages, we can examine the number of revisions that were stored for each document. As mentioned above, each typing break of 300ms or more produced a new revision. A longer document will naturally result in more revisions; thus, the number of revisions normalized by document length is interesting as well.

Table 5.6 shows the distribution of revision counts in absolute numbers, as well as normalized by each document's number of sentences. The exceptionally low minimum number of absolute revisions seen in Table 5.6a can be explained by the fact that very few sources were available for some of the Batch 1 topics. On the other hand, the lower average number of revisions per sentence shown in Table 5.6b may be evidence that authors employ a different editing style—perhaps they are more prone to verbatim copying when given predetermined sources.

To test if the latter number of revisions per sentence correlates with the batch assignment—and thus with the retrieval model used to find sources—we again per-

Table 5.6: Distribution of revision counts for the 297 corpus documents.

(a) Absolute

	Revisions		
	Overall	Batch 1	Batch 2
mean	2 132.4	1 400.4	2 849.8
std	1 447.3	1 049.7	1 426.5
min	45.0	45.0	260.0
median	1 923.0	1 141.0	2 843.0
max	6 975.0	4 651.0	6 975.0

(b) Per sentence

	Revisions / Sentences		
	Overall	Batch 1	Batch 2
mean	9.10	6.34	11.80
std	5.94	4.27	6.11
min	0.47	0.47	1.45
median	8.20	4.94	11.61
max	31.50	18.49	31.50

form a two-sided t-test for paired samples over the 147 topics in both batches. In this case, it results in a p-value of 3×10^{-14} for the null hypothesis, or near certainty for the hypothesis that the expected mean number of revisions differs significantly between batches.

While this is not final evidence of different approaches to plagiarism being prevalent from one batch to the other, it does provide some motivation for exploring this relationship further in future research.

Figure 5.3 shows considerable spread in the number of revisions per sentence between different authors, as well as between different documents written by the same author. While for some documents authors logged as many as thirty revisions per sentence, most achieve a median in the single digits. The number of revisions per sentence may be of interest as a measure of the degree of paraphrasing, in addition to the one explored in Chapter 3.

While there is no doubt that there is much more to study about the documents in Webis-TRC-12 and their revision histories, not all of it can be discussed in the framework of this thesis. We next turn our attention once more to the query logs recording Batch 2 authors' interactions with the search engine.

5.3 Click Trails as Implicit Relevance Judgements

Chapter 4 included a detailed discussion of the authors' interaction with the search engine for all of the Batch 2 topics. One aspect of these query logs that we have not discussed so far is the presence of click trails. In most query log datasets, including the AOL query log and the Webis-SMC-12 dataset derived from it, the available information is limited to the queries entered into the search engine and any entries on the result page that the user has clicked.

By contrast, Webis-TRC-12 includes additional information in that all clicks on links in result documents are also recorded. We refer to clicks on search engine results

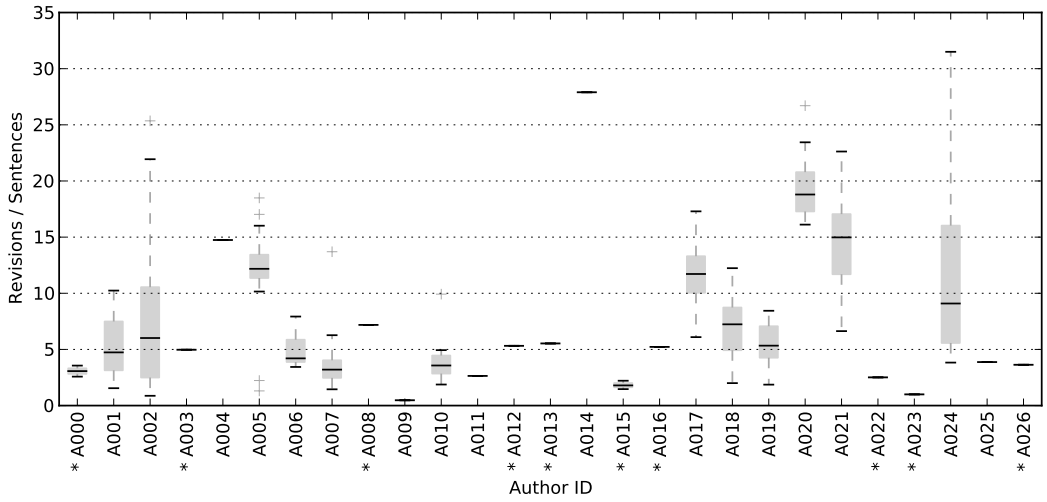


Figure 5.3: Distribution of revisions/sentences by author.

as *search engine result page* (SERP) clicks, and to clicks within result documents as trail clicks. Thus, we can reconstruct a *click trail* to each ClueWeb09 document that users have seen throughout writing their texts. The longest of these click trails spans 51 unique documents, but most of them are very short. In total, only about 15% of all recorded clicks are trail clicks, the rest are SERP clicks.

Since we also know from which sources authors reused text, this represents an important signal regarding the relevance of ClueWeb09 documents to a given topic. For the 150 TREC topics on which Potthast et al. (2012a) based the Webis-TRC-12 topics, relevance judgements from TREC assessors are already available, and were used to compile the list of suggested sources for Batch 1. TREC annotators use a six-point Likert scale to judge the relevance of a document. The scale ranges from "spam" (not relevant to any conceivable topic), to "irrelevant" (to the topic under consideration), on to "relevant" and "key", i.e., highly relevant.

We can infer similar relevance judgements from the browsing and text reuse behavior of our Batch 2 authors. We consider documents used as sources to be of key relevance, documents that lie on a click trail to a source to be relevant, and all other documents that the author examined but did not reuse as irrelevant. Thus, we have two different authorities evaluating the relevance of ClueWeb09 documents for each topic.

Table 5.7 compares TREC judgements and author judgements in a contingency table. Each row of the table contains a different relevance judgement of the TREC assessors, and each column a different judgement by our authors. To arrive at the value in a given table cell, we compiled the sets of ClueWeb09 documents judged with the corresponding pair of relevance scores, and computed the size of the intersection. One such intersecting set of corresponding relevance judgements exists for

Table 5.7: Contingency table of TREC judgements versus author judgments.

TREC	**Author judgement**			
judgement	irrelevant	relevant	key	unjudged
spam (-2)	3	0	1	2 446
spam (-1)	64	4	18	16 657
irrelevant (0)	219	13	73	33 567
relevant (1)	114	8	91	10 676
relevant (2)	44	5	56	3 711
key (3)	12	0	8	526
unjudged	5 506	221	1 690	–

each topic—in the table, we report the sums over all topics. The last column of the table contains the sets of documents that were evaluated by TREC judges, but not found by our authors. Conversely, the last row reports on documents retrieved by our authors that had not been previously evaluated by TREC assessors.

As is apparent from the table, there is mixed agreement between the two sources of judgements. Our authors reused even some documents that received a "spam" rating from TREC judges, while dismissing some documents considered of key importance according to their TREC rating. Overall though, the degree of overlap between the sets of documents appearing in the two relevance ratings is very small—most of the documents our authors found using the ChatNoir search engine had previously not been rated in any TREC competition.

This is an important result in itself, as it implies once again that the different retrieval model across both batches did have a significant effect on source selection. For most of the topics in the corpus, the document in Batch 1 must hence be composed of a very different set of sources than the one in Batch 2. As a final note, the table shows that the set of relevant documents as judged by our authors is very small compared to the sets of irrelevant and key documents. This is due to the fact that most of our authors' clicks were on search result pages, and only very few occurred within documents. Under our judgement scheme, a search result click that is not the start of a click trail can only result in a relevant (if the result document is used as a source) or irrelevant judgement (if it is not).

Despite these caveats, we believe that author relevance judgements will prove to be a useful complement to the evaluation done by TREC assessors. This concludes our discussion of relevance judgements in the query log. To complete this chapter, we next discuss some of the software developed in trying to make the data in the corpus more accessible.

5.4 Interactive Corpus Tools

In order to facilitate exploring the data, and to help present it to interested parties, we developed a number of software tools. These tools not only help make sense of the data in the corpus at hand, they may also support quality control work for future crowdsourcing efforts, since they allow us to retrace the writers' work interactively. This section presents some of the software produced to this end.

Software Framework

The initial data collection effort for Webis-TRC-12 resulted in a wide variety of information stored in different forms. This includes the fine-grained document editing history stored in a Git repository, time sheets and billing information provided by the crowdsourcing platform, as well as various search engine and web server log files documenting the authors' search behavior. In order to make these data accessible at interactive speeds—as well as to support the data mining efforts described in preceding chapters—a software framework was developed to provide a unified interface to all of the data.

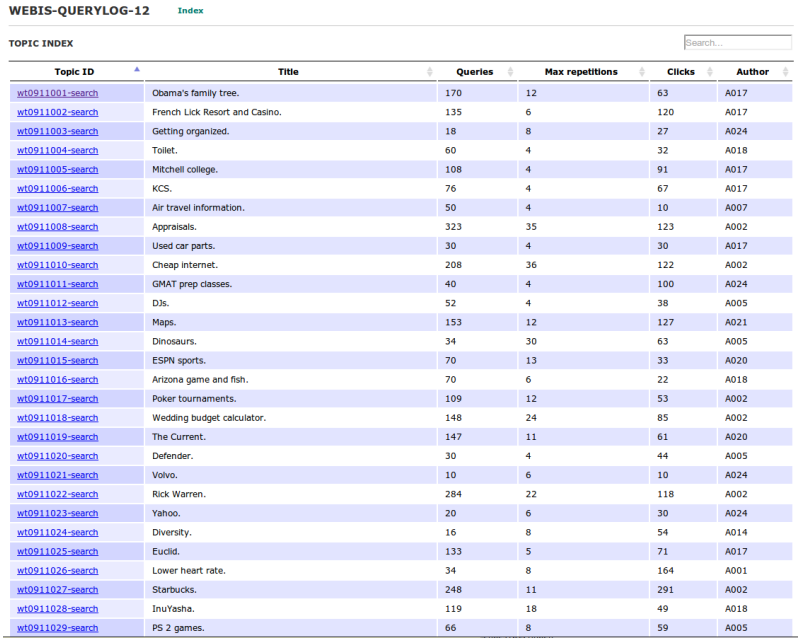
Since parsing and preprocessing all of the various data sources is a somewhat slow process, we store the resulting intermediate data—including individual corpus document revisions and search engine log events—in a relational database. By means of log time stamps and related information, this allows us to correlate data from different sources, and enables even deeper insights into the data in the future. We use an object-relational framework to facilitate programmatic access to the database contents. Since most of the results presented in preceding chapters are based on data analysis done in the Python[2] programming language, we chose the SQLAlchemy toolkit[3] for this function.

As an added bonus, this choice permits us to implement web interfaces for data exploration with minimal extra work. A few of them are described in the remainder of this section.

Webis Querylog Browser

The first data exploration interface intends to help explore the query logs showcased primarily in Chapter 4. Much of the preparatory work for the aforementioned chapter was based on the search for interesting patterns in the query log, which then informed hypotheses concerning their potential causes. This type of work involves

[2] `http://python.org/` (last accessed March 2013)
[3] `http://www.sqlalchemy.org/` (last accessed March 2013)



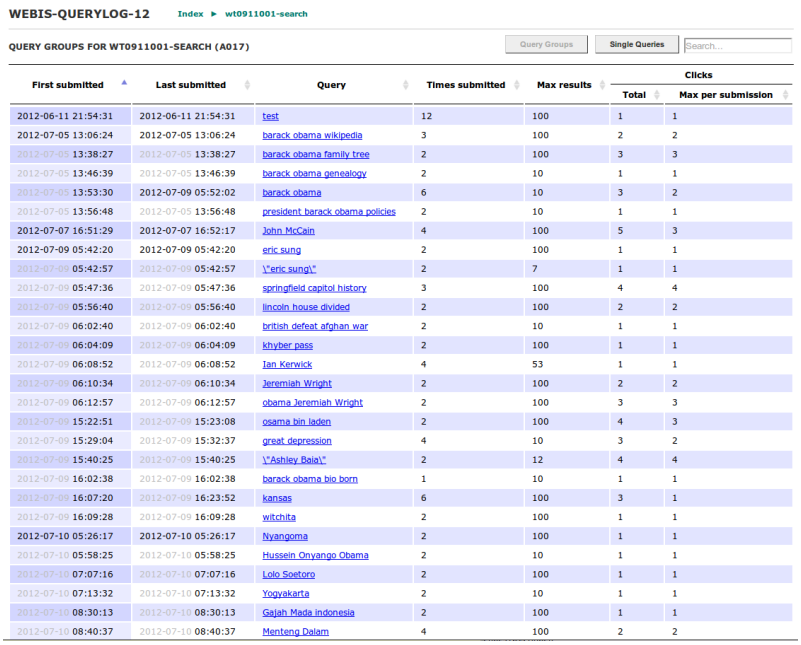
WEBIS-QUERYLOG-12 Index

TOPIC INDEX

Search...

Topic ID	Title	Queries	Max repetitions	Clicks	Author
wt0911001-search	Obama's family tree.	170	12	63	A017
wt0911002-search	French Lick Resort and Casino.	135	6	120	A017
wt0911003-search	Getting organized.	18	8	27	A024
wt0911004-search	Toilet.	60	4	32	A018
wt0911005-search	Mitchell college.	108	4	91	A017
wt0911006-search	KCS.	76	4	67	A017
wt0911007-search	Air travel information.	50	4	10	A007
wt0911008-search	Appraisals.	323	35	123	A002
wt0911009-search	Used car parts.	30	4	30	A017
wt0911010-search	Cheap internet.	208	36	122	A002
wt0911011-search	GMAT prep classes.	40	4	100	A024
wt0911012-search	DJs.	52	4	38	A005
wt0911013-search	Maps.	153	12	127	A021
wt0911014-search	Dinosaurs.	34	30	63	A005
wt0911015-search	ESPN sports.	70	13	33	A020
wt0911016-search	Arizona game and fish.	70	6	22	A018
wt0911017-search	Poker tournaments.	109	12	53	A002
wt0911018-search	Wedding budget calculator.	148	24	85	A002
wt0911019-search	The Current.	147	11	61	A020
wt0911020-search	Defender.	30	4	44	A005
wt0911021-search	Volvo.	10	6	10	A024
wt0911022-search	Rick Warren.	284	22	118	A002
wt0911023-search	Yahoo.	20	6	30	A024
wt0911024-search	Diversity.	16	8	54	A014
wt0911025-search	Euclid.	133	5	71	A017
wt0911026-search	Lower heart rate.	34	8	164	A001
wt0911027-search	Starbucks.	248	11	291	A002
wt0911028-search	InuYasha.	119	18	49	A018
wt0911029-search	PS 2 games.	66	8	59	A005

(a) Topic index



WEBIS-QUERYLOG-12 Index ▸ wt0911001-search

QUERY GROUPS FOR WT0911001-SEARCH (A017)

Query Groups | Single Queries | Search...

First submitted	Last submitted	Query	Times submitted	Max results	Clicks	
					Total	Max per submission
2012-06-11 21:54:31	2012-06-11 21:54:31	test	12	100	1	1
2012-07-05 13:06:24	2012-07-05 13:06:24	barack obama wikipedia	3	100	2	2
2012-07-05 13:38:27	2012-07-05 13:38:27	barack obama family tree	2	100	3	3
2012-07-05 13:46:39	2012-07-05 13:46:39	barack obama genealogy	2	10	1	1
2012-07-05 13:53:30	2012-07-09 05:52:02	barack obama	6	10	3	2
2012-07-05 13:56:48	2012-07-05 13:56:48	president barack obama policies	2	10	1	1
2012-07-07 16:51:29	2012-07-07 16:52:17	John McCain	4	100	5	3
2012-07-09 05:42:20	2012-07-09 05:42:20	eric sung	2	100	1	1
2012-07-09 05:42:57	2012-07-09 05:42:57	\"eric sung\"	2	7	1	1
2012-07-09 05:47:36	2012-07-09 05:47:36	springfield capitol history	3	100	4	4
2012-07-09 05:56:40	2012-07-09 05:56:40	lincoln house divided	2	100	2	2
2012-07-09 06:02:40	2012-07-09 06:02:40	british defeat afghan war	2	10	1	1
2012-07-09 06:04:09	2012-07-09 06:04:09	khyber pass	2	100	1	1
2012-07-09 06:08:52	2012-07-09 06:08:52	Ian Kerwick	4	53	1	1
2012-07-09 06:10:34	2012-07-09 06:10:34	Jeremiah Wright	2	100	2	2
2012-07-09 06:12:57	2012-07-09 06:12:57	obama Jeremiah Wright	2	100	3	3
2012-07-09 15:22:51	2012-07-09 15:23:08	osama bin laden	2	100	4	3
2012-07-09 15:29:04	2012-07-09 15:32:37	great depression	4	10	3	2
2012-07-09 15:40:25	2012-07-09 15:40:25	\"Ashley Baia\"	2	12	4	4
2012-07-09 16:02:38	2012-07-09 16:02:38	barack obama bio born	1	10	1	1
2012-07-09 16:07:20	2012-07-09 16:23:52	kansas	6	100	3	1
2012-07-09 16:09:28	2012-07-09 16:09:28	witchita	2	100	1	1
2012-07-10 05:26:17	2012-07-10 05:26:17	Nyangoma	2	100	1	1
2012-07-10 05:58:25	2012-07-10 05:58:25	Hussein Onyango Obama	2	10	1	1
2012-07-10 07:07:16	2012-07-10 07:07:16	Lolo Soetoro	2	100	1	1
2012-07-10 07:13:32	2012-07-10 07:13:32	Yogyakarta	2	10	1	1
2012-07-10 08:30:13	2012-07-10 08:30:13	Gajah Mada indonesia	2	100	1	1
2012-07-10 08:40:37	2012-07-10 08:40:37	Menteng Dalam	4	100	2	2

(b) Query log

Figure 5.4: User interface for the Webis Querylog Browser.

viewing the data from many different angles, a task which becomes very cumbersome when based on plain-text server logs, or even a direct interface to a relational database.

Instead, we chose to implement a corpus browsing interface as a dynamic web page that leverages the full capabilities of hypertext markup and scripting languages to help explore the data from different directions. As appropriate for a web-based solution, our implementation follows a client-server architecture. On the server side, we implemented a web service according to the well-established Representational State Transfer (REST) architectural pattern introduced by Fielding (2000). This service acts as a gateway between database and user interface, and translates the raw data into an easily-presentable form.

On the client side, we implemented a user interface using HTML and JavaScript, based in large part on the jQuery DataTables library.[4] This software framework supports our corpus browsing interface, and allows sorting and filtering the data by arbitrary criteria.

Figure 5.4 shows two screenshots of the querylog browser user interface. The first presents the topic index, which shows a high-level overview of all the data in the corpus. For every topic, it shows the number of clicks and queries, the maximum number of times a unique query was repeated, as well as the author. Using the sorting and filtering tools, we can identify topics with interesting interaction levels, or compare the interaction logs for a single author on a high level. Clicking on a topic ID switches to a more detailed view of the corresponding interaction log.

The second screenshot shows a view of one topic's query stream. As shown, it displays cumulative statistics for each distinct query string, including the dates of first and last occurrence, and the cumulative number of clicks on results of the query. Aside from this aggregate perspective, a view of the raw query stream in the sequence it occurred is also available. Queries can be sorted by date of submission, or by the amount of related click interaction. Other views in the querylog browser allow drilling down to the level of individual clicks or search result pages.

As an added benefit to supporting our own research, this web-based frontend to the corpus permits easy sharing of the data with interested third parties. By hosting such browsing interfaces on a public-facing web server, we can allow others to evaluate the suitability of our data for their purposes.

Edit History Viewer

The other main component of the corpus—the documents themselves and their revisions—lends itself to a different approach to interactive exploration. While the history flow visualization introduced in Section 3.4 give a good high-level view of

[4] `http://www.datatables.net/` (last accessed March 2013)

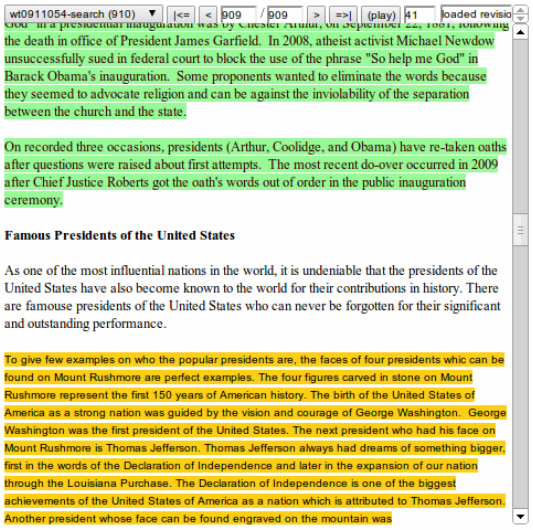


Figure 5.5: Edit history viewer showing the final revision for topic 054 from Batch 2.

how the length and composition of a document changes over time, they do not allow us to examine different versions of the text itself. Ideally, we want to be able to read each individual revision of a document so that we can retrace the steps a given author took in the process of compiling the final text.

For this purpose, we implemented another client-server web interface to the data. A REST-based server backend that answers requests for individual document revisions stored in the database is complemented by an HTML/JavaScript client that shows the current document revision in a view similar to the text editor used by corpus authors. The client UI, shown in Figure 5.5, includes controls for stepping forwards and backwards through the revision history, and for playing back the entire revision history of the document as in a time-lapse movie.

In order to better follow the author's edits, the server component inserts an HTML anchor element at the position of the most recent edit when serving revisions. The client scrolls the document view to this element whenever it loads a new revision.

While the tools presented in this section have proven invaluable to our understanding of the data in the corpus, there is much room for further development. For future work, we are evaluating the possibility of implementing a unified corpus browser that would tie all of the separate data sources together. For instance, the two tools shown above could be augmented by displaying search engine queries immediately preceding the current edit in the edit history viewer, or by including links in the querylog browser to the revision that was current at the time of a given query.

5.5 Summary

In this chapter, we have explored some of the logistics of crowdsourcing a large text reuse corpus that are in evidence in the recorded data, discussed some additional insights not found in previous chapters, and showcased some of our more practical results. In Section 5.1, we have shone some additional light on the author dimension of the dataset, analyzed how the work of corpus authorship was distributed among different people, and given an impression of our authors' identity and experience as reflected in the questionnaires they completed.

We then discussed how the length of documents is distributed across the corpus' batch division, and explored how our authors' engagement of sources through the search engine implies a relevance judgement of potential source documents. Finally, we gave a brief glimpse into the software development aspects of the work at hand, and showcased some of our work that may be of use to similar efforts in the future.

6 Conclusions

We have discussed various aspects of our investigation of the crowdsourced, simulated plagiarism in the Webis-TRC-12 dataset in this thesis. In the process of our research, we consolidated a large, diverse set of different data sources and made them accessible to further data mining. Drawing on the diverse types of data available, we developed a new approach to categorizing the kinds of text reuse found in a dataset of this kind, and demonstrated the relationship of the search engine interactions recorded in the Webis-TRC-12 to those in a reference dataset. In the present chapter, we summarize our research and its main findings, and point out possible avenues for future research.

Areas of Investigation and Main Findings

At the start of this thesis, we formulated two research questions that we proceeded to address using the data in the Webis-TRC-12. One of them was the question of how previous efforts at categorizing plagiarism could be applied to a crowdsourced corpus. In our survey of two past systems of categorization in Section 2.6, we found that most of the categories they define make little sense in the context of a simulated plagiarism dataset. In Chapter 3, we proceeded to synthesize a subset of the categories into a plagiarism spectrum of our own. Its dimensions—the degrees of paraphrasing and interleaving found in a corpus document—distinguish documents by their treatment of their sources.

We defined a set of empirical measures to locate our corpus documents within this space, and found that we can not only usefully separate the documents, but also find some reflection of the author's personal style. Involving an additional dimension of the data—the way documents change over time—we found evidence of two fundamental strategies for editing a corpus document: build-up and boil-down. While not all corpus authors can be clearly identified as favoring either a build-up or a boil-down approach to text reuse, we found that especially the former seem to stick to their preferred strategy.

In Chapter 4, we explored authors' interaction with the search engine, using the Webis-SMC-12 dataset introduced in Section 2.5 as a reference corpus for comparison. We found that while the search missions in the Webis-TRC-12 query logs are indeed of novel quality, users of public web search engines do tackle the kinds of exploratory search tasks that our authors pursue. Via the author dimension,

we managed to bridge the gap to the editing strategies investigated in Chapter 3. While this area will surely require further investigation, we found some evidence of a correlation between editing strategies and the time distribution of search engine interactions in the query log.

Aside from the fact that without writers, compiling a corpus of writing samples is impossible, the preferences and strategies of different writers seem to tie much of our disparate research together. With this in mind, we dedicated the first part of Chapter 5 to the crowdsourcing effort. After exploring the distribution of work, the authors' demographics and experiences, and the resource investment involved in a corpus construction effort of this magnitude, we investigated the influence of the way authors access their sources on basic statistics of the corpus documents. We found statistically significant differences between a first batch of documents compiled from a limited set of predetermined sources, and a second batch where authors had access to a full search engine.

We concluded Chapter 5 by showcasing a pair of software tools we developed to make exploring the data in the corpus more interactive and insightful. Even though we have gained many new insights from the data already, we believe the potential of Webis-TRC-12 is far from exhausted.

Future Work

In Chapter 1, we named three research communities that we feel will benefit the most from the data in the Webis-TRC-12: plagiarism, search, and paraphrasing. Throughout our investigation, the former two have received most of the attention, whereas paraphrasing was covered in much less detail. Part of the reason for this is the fact that the low-hanging fruit in the data benefit mostly plagiarism and search. We could almost immediately make use of the final revisions of the corpus documents to study and organize the kinds of plagiarism found in the data. The search engine logs required only a small amount of post-processing before we were able to extract features and compare them to a reference dataset.

In order to extract useful paraphrasing information, we must delve deeper into the data, but we nevertheless believe that future paraphrasing research has much to gain from Webis-TRC-12. For instance, by identifying individual passages copied from source documents, and then tracking how they change throughout the revisions made to corpus documents, we could construct a novel quality of paraphrasing dataset. Current paraphrasing corpora usually consist only of pairs of text units (such as sentences or paragraphs) where one is a paraphrase of the other. Using the information in Webis-TRC-12, we may be able to fully model the process from original passage to modified passage as a sequence of paraphrasing operations.

There is an additional aspect to the time dimension of corpus documents that we have not investigated: how a single author's text reuse strategy evolves across multiple documents. Some of the authors spent many weeks working on the corpus, writing ten or more documents in the process. It is conceivable that, as they become more accustomed to reusing text, the way they plagiarize changes over time.

In a similar vein, it may be of interest to further compare the behavior of paid authors versus volunteers. Writers hired through the crowdsourcing platform were paid by the hour. As such, they have a financial incentive to spend as much time as possible working on a single document. Volunteers, on the other hand, would be more interested in getting the job done quickly and getting on with their lives—arguably, a situation more in line with that of a real plagiarist. An investigation of this aspect may lead to advances in task design that permit an even more realistic modeling of plagiarism in future corpora, for instance by offering writers a flat fee per document instead of an hourly salary. However, since less than 5% of the documents in Webis-TRC-12 were actually written by volunteers, it may prove difficult to obtain statistically significant results.

While we have investigated diverse aspects of the data, many possible directions for future research remain. The ones we have sketched above only scratch the surface of the potential of this dataset.

Bibliography

Eugene Agichtein, Ryen W. White, Susan T. Dumais, and Paul N. Bennett. Search, interrupted: Understanding and predicting search task continuation. In *SIGIR*, pages 315–324, 2012.

Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

Steven Burrows, Martin Potthast, and Benno Stein. Paraphrase acquisition via crowdsourcing and machine learning. *Transactions on Intelligent Systems and Technology (ACM TIST)*, 2012.

Vitor R. Carvalho, Matthew Lease, and Emine Yilmaz. Crowdsourcing for search evaluation. *SIGIR Forum*, 44(2):17–22, 2011.

Paul Clough and Mark Stevenson. Developing a corpus of plagiarised short answers. *Language Resources and Evaluation*, 45(1):5–24, 2011.

Paul Clough, Robert Gaizauskas, and Scott Piao. Building and annotating a corpus for the study of journalistic text reuse. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-02)*, volume 5, pages 1678–1691, 2002.

Brian S. Everitt and Anders Skrondal. *The Cambridge Dictionary of Statistics*. Cambridge University Press, 4th edition, 2010.

Roy Thomas Fielding. *Architectural styles and the design of network-based software architectures*. PhD thesis, University of California, 2000.

Daniel Gayo-Avello. A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences*, 179(12):1822–1843, 2009.

Matthias Hagen, Jakob Gomoll, and Benno Stein. Improved cascade for search mission detection. In *ECIR 12 Workshop on Information Retrieval over Query Sessions (SIR 12)*, 2012. URL `http://ir.cis.udel.edu/ECIR12Sessions/`. (last accessed Mar 2013).

Matthias Hagen, Jakob Gomoll, and Benno Stein. From search session detection to search mission detection. In *Proceedings of the 10th International Conference Open Research Areas in Information Retrieval (OAIR 13) (to appear)*. ACM, 2013.

Rosie Jones and Kristina L. Klinkner. Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 699–708. ACM, 2008.

Melanie Kellar, Carolyn Watters, and Michael Shepherd. A field study characterizing web-based information-seeking tasks. *Journal of the American Society for Information Science and Technology*, 58(7):999–1018, 2007.

Alexander Kotov, Paul N. Bennett, Ryen W. White, Susan T. Dumais, and Jaime Teevan. Modeling and analysis of cross-session search tasks. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 5–14. ACM, 2011.

Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. Identifying task-based sessions in search engine query logs. In *Proceedings of the 4th ACM international conference on Web search and data mining*, pages 277–286. ACM, 2011.

Gary Marchionini. Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.

Mark Nolan. IA column: Exploring exploratory search. *Bulletin of the American Society for Information Science and Technology*, 34(4):38–41, 2008.

Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. In *Proceedings of the 1st international conference on scalable information systems*, page 1. ACM, 2006.

Martin Potthast. Crowdsourcing a wikipedia vandalism corpus. In *Proceedings of the 33rd International ACM Conference on Research and Development in Information Retrieval (SIGIR 10)*, pages 789–790. ACM, 2010.

Martin Potthast. *Technologies for Reusing Text from the Web.* Dissertation, Bauhaus-Universität Weimar, 2011.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. An evaluation framework for plagiarism detection. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, 2010.

Martin Potthast, Andreas Eiselt, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. Overview of the 3rd international competition on plagiarism detection. In *Working Notes Papers of the CLEF 2011 Evaluation Labs*, 2011.

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 *Working Notes Papers of the CLEF 2012 Evaluation Labs and Workshop*, 2012a.

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 Conference on Research and Development in Information Retrieval (SIGIR 12)*, page 1004. ACM, 2012b.

Abigail J. Sellen, Rachel Murphy, and Kate L. Shaw. How knowledge workers use the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, pages 227–234. ACM, 2002.

Turnitin. *White Paper – The Plagiarism Spectrum*. iParadigms, LLC, 2012. URL `http://turnitin.com/assets/en_us/media/plagiarism_spectrum.php`. (last accessed Mar 2013).

Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 575–582. ACM, 2004.

VroniPlag. Categories of plagiarism (in german), 2012. URL `http://de.vroniplag.wikia.com/wiki/VroniPlag_Wiki:Grundlagen/PlagiatsKategorien?oldid=112731`. (last accessed Mar 2013).

Ryen W White, Bill Kules, Steven M Drucker, et al. Supporting exploratory search, introduction, special issue, communications of the acm. *Communications of the ACM*, 49(4):36–39, 2006.

Ryen W White, Gary Marchionini, and Gheorghe Muresan. Editorial: Evaluating exploratory search systems. *Information Processing Management*, 44(2):433–436, 2008.

Kelly H Zou, Kemal Tuncali, and Stuart G Silverman. Correlation and simple linear regression. *Radiology*, 227(3):617–628, 2003.