Exploratory Search Missions for TREC Topics

Martin Potthast Matthias Hagen Michael Völske Benno Stein

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

ABSTRACT

We report on the construction of a new query log corpus that consists of 150 exploratory search missions, each of which corresponds to one of the topics used at the TREC Web Tracks 2009–2011. Involved in the construction was a group of 12 professional writers, hired at the crowdsourcing platform oDesk, who were given the task to write essays of 5000 words length about these topics, thereby inducing genuine information needs. The writers used a ClueWeb09 search engine for their research to ensure reproducibility. Thousands of queries, clicks, and relevance judgments were recorded. This paper overviews the research that preceded our endeavors, details the corpus construction, gives quantitative and qualitative analyses of the data obtained, and provides original insights into the querying behavior of writers. With our work we contribute a missing building block in a relevant evaluation setting in order to allow for better answers to questions such as: "What is the performance of today's search engines on exploratory search?" and "How can it be improved?" The corpus will be made publicly available.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Query formulation

Keywords: Query Log, Exploratory Search, Search Missions

1. INTRODUCTION

Humans frequently conduct task-based information search, i.e., they interact with search appliances in order to conduct the research deemed necessary to solve knowledge-intensive tasks. Examples include long-lasting interactions which may involve many search sessions spread out across several days. Modern web search engines, however, are optimized for the diametrically opposed task, namely to answer short-term, atomic information needs. Nevertheless, research has picked up this challenge: in recent years, a number of new solutions for exploratory search have been proposed and evaluated. However, most of them involve an overhauling of the entire search experience. We argue that exploratory search tasks are already being tackled, after all, and that this fact has not been sufficiently investigated. Reasons for this shortcoming can be found in the lack of publicly available data to be studied. Ideally, for any given task that fits the aforementioned description, one would have a large set of search interaction logs from a diversity of humans solving it. Obtaining such data, even for a single task, has not been done at scale until now. Even search companies, which have access to substantial amounts of raw query log data, face difficulties in discerning individual exploratory tasks from their logs.

In this paper, we contribute by introducing the first large corpus of long, exploratory search missions. The corpus was constructed via crowdsourcing by employing writers whose task was to write long essays on given TREC topics, using a ClueWeb09 search engine for research. Hence, our corpus forms a strong connection to existing evaluation resources that are used frequently in information retrieval. Further, it captures the way how average users perform exploratory search today, using state-of-the-art search interfaces. The new corpus is intended to serve as a point of reference for modeling users and tasks as well as for comparison with new retrieval models and interfaces. Key figures of the corpus are shown in Table 2.

After a brief review of related work, Section 2 details the corpus construction and Section 3 gives first quantitative and qualitative analyses, concluding with insights into writers' search behavior.

1.1 Related Work

To date, the most comprehensive overview of research on exploratory search systems is that of White and Roth [19]. More recent contributions not covered in this body of work include the approaches proposed by Morris et al. [13], Bozzon et al. [2], Cartright et al. [4], and Bron et al. [3]. Exploratory search is studied also within contextual IR and interactive IR, as well as across disciplines, including human computer interaction, information visualization, and knowledge management.

Regarding the evaluation of exploratory search systems, White and Roth [19] conclude that "traditional measures of IR performance based on retrieval accuracy may be inappropriate for the evaluation of these systems" and that "exploratory search evaluation [...] must include a mixture of naturalistic longitudinal studies" while "[...] simulations developed based on interaction logs may serve as a compromise between existing IR evaluation paradigms and [...] exploratory search evaluation." The necessity of user studies makes evaluations cumbersome and, above all, expensive. By providing part of the solution (a decent corpus) for free, we want to overcome the outlined difficulties. Our corpus compiles a solid database of exploratory search behavior, which researchers may use for comparison purposes as well as for bootstrapping simulations.

Regarding standardized resources to evaluate exploratory search, hardly any have been published up to now. White et al. [18] dedicated a workshop to evaluating exploratory search systems in which requirements, methodologies, as well as some tools have been proposed. Yet, later on, White and Roth [19] found out that still no "methodological rigor" has been reached—a situation which has not changed much until today. The departure from traditional evaluation methodologies (such as the Cranfield paradigm) and resources (especially those employed at TREC) has lead researchers to devise ad-hoc evaluations which are mostly incomparable across papers and which cannot be reproduced easily.

A potential source of data for the purpose of assessing current exploratory search behavior is to detect exploratory search tasks within raw search engine logs, such as the 2006 AOL query log [14].

However, most session detection algorithms deal with short term tasks only and the few algorithms that aim to detect longer search missions still have problems when detecting interesting semantic connections of intertwined search tasks [10, 12, 8]. In this regard, our corpus may be considered the first of its kind.

To justify our choice of an exploratory task, namely that of writing an essay about a given TREC topic, we refer to Kules and Capra [11], who manually identified exploratory tasks from raw query logs on a small scale, most of which turned out to involve writing on a given subject. Egusa et al. [6] describe a user study in which they asked participants to do research for a writing task, however, without actually writing something. This study is perhaps closest to ours, although the underlying data has not been published. The most notable distinction is that we asked our writers to actually write, thereby creating a much more realistic and demanding state of mind since their essays had to be delivered on time.

2. CORPUS CONSTRUCTION

As discussed in the related work, essay writing is considered a valid approach to study exploratory search. Two data sets form the basis for constructing a respective corpus, namely (1) a set of topics to write about and (2) a set of web pages to research about a given topic. With regard to the former, we resort to topics used at TREC, specifically to those from the Web Tracks 2009–2011. With regard to the latter, we employ the ClueWeb09 (and not the "real web in the wild"). The ClueWeb09 consists of more than one billion documents from ten languages; it comprises a representative cross-section of the real web, is a widely accepted resource among researchers, and it is used to evaluate the retrieval performance of search engines within several TREC tracks. The connection to TREC will strengthen the compatibility with existing evaluation methodology and allow for unforeseen synergies. Based on the above decisions, our corpus construction steps can be summarized as follows:

1. Rephrasing of the 150 topics used at the TREC Web Tracks 2009–2011 so that they invite people to write an essay.
2. Indexing of the English portion of the ClueWeb09 (about 0.5 billion documents) using the BM25F retrieval model plus additional features.
3. Development of a search interface that allows for answering queries within milliseconds and that is designed along the lines of commercial search interfaces.
4. Development of a browsing interface for the ClueWeb09, which serves ClueWeb09 pages on demand and which rewrites links on delivered pages so that they point to their corresponding ClueWeb09 pages on our servers.
5. Recruiting 12 professional writers at the crowdsourcing platform oDesk from a wide range of hourly rates for diversity.
6. Instructing the writers to write essays of at least 5000 words length (corresponds to an average student's homework assignment) about an open topic among the initial 150, using our search engine and browsing only ClueWeb09 pages.
7. Logging all writers' interactions with the search engine and the ClueWeb09 on a per-topic basis at our site.
8. Double-checking all of the 150 essays for quality.

After the deployment of the search engine and successfully completed usability tests (see Steps 2-4 and 7 above), the actual corpus construction took nine months, from April 2012 through December 2012. The post-processing of the data took another four months, so that this corpus is among the first, late-breaking results from our efforts. However, the outlined experimental setup can obviously serve different lines of research. The remainder of the section presents elements of our setup in greater detail.

Used TREC Topics.

Since the topics from the TREC Web Tracks 2009–2011 were not amenable for our purpose as is, we rephrased them so that they ask for writing an essay instead of searching for facts. Consider for example topic 001 from the TREC Web Track 2009:

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

This topic is rephrased as follows:

> *Obama's family*. Write about President Barack Obama's family history, including genealogy, national origins, places and dates of birth, etc. Where did Barack Obama's parents and grandparents come from? Also include a brief biography of Obama's mother.

In the example, Sub-topic 1 is considered too specific for our purposes while the other sub-topics are retained. TREC Web track topics divide into faceted and ambiguous topics. While topics of the first kind can be directly rephrased into essay topics, from topics of the second kind one of the available interpretations is chosen.

A Search Engine for Controlled Experiments.

To give the oDesk writers a familiar search experience while maintaining reproducibility at the same time, we developed a tailored search engine called ChatNoir [15]. Besides ours, the only other public search engine for the ClueWeb09 is hosted at Carnegie Mellon and based on Indri. Unfortunately, it is far from our efficiency requirements. Our search engine returns results after a couple of hundreds of milliseconds, its interface follows industry standards, and it features an API that allows for user tracking.

ChatNoir is based on the BM25F retrieval model [17], uses the anchor text list provided by Hiemstra and Hauff [9], the PageRanks provided by the Carnegie Mellon University,[1] and the spam rank list provided by Cormack et al. [5]. ChatNoir comes with a proximity feature with variable-width buckets as described by Elsayed et al. [7]. Our choice of retrieval model and ranking features is intended to provide a reasonable baseline performance. However, it is neither near as mature as those of commercial search engines nor does it compete with the best-performing models proposed at TREC. Yet, it is among the most widely accepted models in the information retrieval community, which underlines our goal of reproducibility.

In addition to its retrieval model, ChatNoir implements two search facets: text readability scoring and long text search. The former facet, similar to that provided by Google, scores the readability of a text found on a web page via the well-known Flesh-Kincaid grade level formula: it estimates the number of years of education required in order to understand a given text. This number is mapped onto the three categories "simple", "intermediate", and "expert." The long text search facet omits search results which do not contain at least one continuous paragraph of text that exceeds 300 words. The two facets can be combined with each other. They are meant to support writers that want to reuse text from retrieved search results. Especially interesting for this type of writers are result documents containing longer text passages and documents of a specific reading

[1] http://boston.lti.cs.cmu.edu/clueWeb09/wiki/tiki-index.php?page=PageRank

Table 1: Demographics of the twelve writers employed.

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

level such that reusing text from the results still yields an essay with homogeneous readability.

When clicking on a search result, ChatNoir does not link into the real web but redirects into the ClueWeb09. Though ClueWeb09 provides the original URLs from which the web pages have been obtained, many of these page may have gone or been updated since. We hence set up an interface that serves web pages from the ClueWeb09 on demand: when accessing a web page, it is pre-processed before being shipped, removing all kinds of automatic referrers and replacing all links to the real web with links to their counterpart inside ClueWeb09. This way, the ClueWeb09 can be browsed as if surfing the real web and it becomes possible to track a user's movements. The ClueWeb09 is stored in the HDFS of our 40 node Hadoop cluster, and web pages are fetched with latencies of about 200ms. ChatNoir's inverted index has been optimized to guarantee fast response times, and it is deployed on the same cluster.

Hired Writers.

Our ideal writer has experience in writing, is capable of writing about a diversity of topics, can complete a text in a timely manner, possesses decent English writing skills, and is well-versed in using the aforementioned technologies. This wish list lead us to favor (semi-)professional writers over, for instance, volunteer students recruited at our university. To hire writers, we made use of the crowdsourcing platform oDesk.[2] Crowdsourcing has quickly become one of the cornerstones for constructing evaluation corpora, which is especially true for paid crowdsourcing. Compared to Amazon's Mechanical Turk [1], which is used more frequently than oDesk, there are virtually no workers at oDesk submitting fake results due to advanced rating features for workers and employers.

Table 1 gives an overview of the demographics of the writers we hired, based on a questionnaire and their resumes at oDesk. Most of them come from an English-speaking country, and almost all of them speak more than one language, which suggests a reasonably good education. Two thirds of the writers are female, and all of them have years of writing experience. Hourly wages were negotiated individually and range from 3 to 34 US-dollars (dependent on skill and country of residence), with an average of about 12 US-dollars. In total, we spent 20 468 US-dollars to pay the writers.

3. CORPUS ANALYSIS

This section presents the results of a preliminary corpus analysis that gives an overview of the data and sheds some light onto the search behavior of writers doing research.

[2] http://www.odesk.com

Table 2: Key figures of our exploratory search mission corpus.

Corpus Characteristic	Distribution min	avg	max	stdev	Σ
Writers					12
Topics					150
Topics / Writer	1	12.5	33	9.3	
Queries					13 651
Queries / Topic	4	91.0	616	83.1	
Clicks					16 739
Clicks / Topic	12	111.6	443	80.3	
Clicks / Query	0	0.8	76	2.2	
Sessions					931
Sessions / Topic	1	12.3	149	18.9	
Days					201
Days / Topic	1	4.9	17	2.7	
Hours					2068
Hours / Writer	3	129.3	679	167.3	
Hours / Topic	3	7.5	10	2.5	
Irrelevant					5962
Irrelevant / Topic	1	39.8	182	28.7	
Irrelevant / Query	0	0.5	60	1.4	
Relevant					251
Relevant / Topic	0	1.7	7	1.5	
Relevant / Query	0	0.0	4	0.2	
Key					1937
Key / Topic	1	12.9	46	7.5	
Key / Query	0	0.2	22	0.7	

Corpus Statistics.

Table 2 shows key figures of the query logs collected, including the absolute numbers of queries, relevance judgments, working days, and working hours, as well as relations among them. On average, each writer wrote 12.5 essays, while two wrote only one, and one very prolific writer managed more than 30 essays.

From the 13 651 submitted queries, each topic got an average of 91. Note that queries often were submitted twice requesting more than ten results or using different facets. Typically, about 1.7 results are clicked for consecutive instances of the same query. For comparison, the average number of clicks per query in the aforementioned AOL query log is 2.0. In this regard, the behavior of our writers on individual queries does not seem to differ much from that of the average AOL user in 2006. Most of the clicks we recorded are search result clicks, whereas 2457 of them are browsing clicks on web page links. Among the browsing clicks, 11.3% are clicks on links that point to the same web page (i.e., anchor links using a URL's hash part). The longest click trail observed lasted 51 unique web pages but most click trails are very short. This is surprising, since we expected a larger proportion of browsing clicks, but it also shows our writers relied heavily on the search engine. If this behavior generalizes, the need for a more advanced support of exploratory search tasks from search engines becomes obvious.

The queries of each writer can be divided into a total of 931 sessions with an average 12.3 sessions per topic. Here, a session is defined as a sequence of queries recorded on a given topic which is not divided by a break longer than 30 minutes. Despite other claims in the literature (e.g., in [10]), we argue that, in our case, sessions can be reliably identified by means of a timeout because of our a priori knowledge about which query belongs to which topic (i.e., task). Typically, finishing an essay took 4.9 days, which fits well the definition of exploratory search tasks being long-lasting.

In their essays, writers referred to web pages they found during their search, citing specific passages and topic-related information used in their texts. This forms an interesting relevance signal which allows us to separate irrelevant from relevant web pages. Slightly different to the terminology of TREC, we consider web pages referred to in an essay as key documents for its respective topic, whereas web pages that are on a click trail leading to a key document are

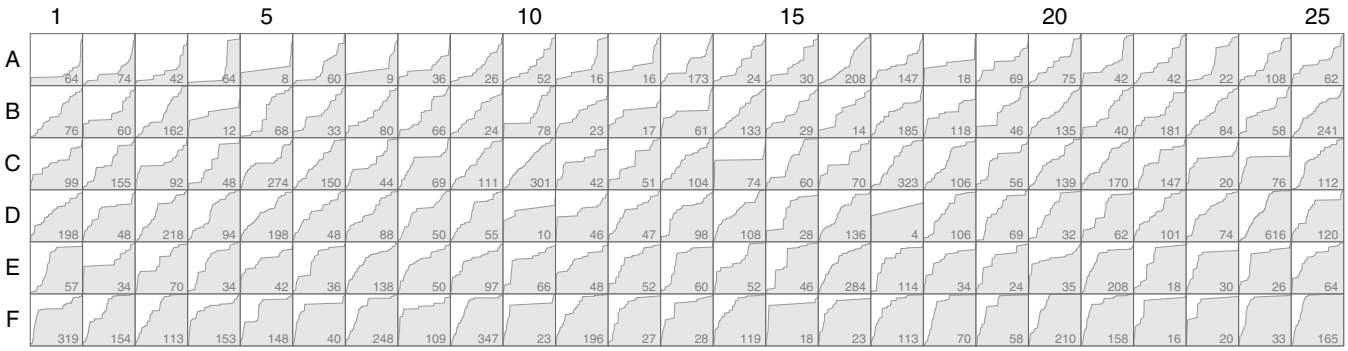

Figure 1: Spectrum of writer search behavior. Each grid cell corresponds to one of the 150 topics and shows a curve of the percentage of submitted queries (y-axis) at times between the first query until the essay was finished (x-axis). The numbers denote the amount of queries submitted. The cells are sorted by area under the curve from the smallest area in cell A1 to the largest area in cell F25.

relevant. The fact, that there are only few click trails of this kind explains the unusually high number of key documents compared to that of relevant ones. The remainder of web pages which were accessed but discarded by our writers may be considered irrelevant.

The writer's search interactions are made freely available as the Webis-Query-Log-12.[3] Note that the writing interactions are the focus of our accompanying ACL paper [16] and contained in the Webis text reuse corpus 2012 (Webis-TRC-12).

Exploring Exploratory Search Missions.

To get an inkling of the wealth of data in our corpus, and how it may influence the design of exploratory search systems, we analyze the writers' search behavior during essay writing. Figure 1 shows for each of the 150 topics a curve of the percentage of queries at any given time between a writer's first query and an essay's completion. We have normalized the time axis and excluded working breaks of more than five minutes. The curves are organized so as to highlight the spectrum of different search behaviors we have observed: in row A, 70–90% of the queries are submitted toward the end of the writing task, whereas in row F almost all queries are submitted at the beginning. In between, however, sets of queries are often submitted in short "bursts," followed by extended periods of writing, which can be inferred from the plateaus in the curves (e.g., cell C12). Only in some cases (e.g., cell C10) a linear increase of queries over time can be observed for a non-trivial amount of queries, which indicates continuous switching between searching and writing.

From these observations, it can be inferred that query frequency alone is not a good indicator of task completion or the current stage of a task, but different algorithms are required for different mission types. Moreover, exploratory search systems have to deal with a broad subset of the spectrum and be able to make the most of few queries, or be prepared that writers interact only a few times with them. Our ongoing research on this aspect focuses on predicting the type of search mission, since we found it does not simply depend on the writer or a topic's difficulty as perceived by the writer.

4. SUMMARY

We introduce the first corpus of search missions for the exploratory task of writing. The corpus is of representative scale, comprising 150 different writing tasks and thousands of queries, clicks, and relevance judgments. A preliminary corpus analysis shows the wide variety of different search behavior to expect from a writer conducting research online. We expect further insights from a forthcoming in-depth analysis, whereas the results mentioned demonstrate the utility of our publicly available corpus.

[3] http://www.webis.de/research/corpora

5. REFERENCES

[1] J. Barr and L. F. Cabrera. AI gets a brain. *Queue*, 4(4):24–29, 2006.

[2] A. Bozzon, M. Brambilla, S. Ceri, and P. Fraternali. Liquid query: multi-domain exploratory search on the web. *Proc. of WWW 2010*.

[3] M. Bron, J. van Gorp, F. Nack, M. de Rijke, A. Vishneuski, and S. de Leeuw. A subjunctive exploratory search interface to support media studies researchers. *Proc. of SIGIR 2012*.

[4] M.-A. Cartright, R. White, and E. Horvitz. Intentions and attention in exploratory health search. *Proc. of SIGIR 2011*.

[5] G. Cormack, M. Smucker, and C. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Information Retrieval*, 14(5):441–465, 2011.

[6] Y. Egusa, H. Saito, M. Takaku, H. Terai, M. Miwa, and N. Kando. Using a concept map to evaluate exploratory search. *Proc. of IIiX 2010*.

[7] T. Elsayed, J. Lin, and D. Metzler. When close enough is good enough: approximate positional indexes for efficient ranked retrieval. *Proc. of CIKM 2011*.

[8] M. Hagen, J. Gommoll, A. Beyer, and B. Stein. From search session detection to search mission detection. *Proc. of SIGIR 2012*.

[9] D. Hiemstra and C. Hauff. MIREX: MapReduce information retrieval experiments. Tech. Rep. TR-CTIT-10-15, University of Twente, 2010.

[10] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. *Proc. of CIKM 2008*.

[11] B. Kules and R. Capra. Creating exploratory tasks for a faceted search interface. *Proc. of HCIR 2008*.

[12] C. Lucchese, S. Orlando, R. Perego, F. Silvestri, and G. Tolomei. Identifying task-based sessions in search engine query logs. *Proc. of WSDM 2011*.

[13] D. Morris, M. Ringel Morris, and G. Venolia. SearchBar: a search-centric web history for task resumption and information re-finding. *Proc. of CHI 2008*.

[14] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. *Proc. of Infoscale 2006*.

[15] M. Potthast, M. Hagen, B. Stein, J. Graßegger, M. Michel, M. Tippmann, and C. Welsch. ChatNoir: a search engine for the ClueWeb09 corpus. *Proc. of SIGIR 2012*.

[16] M. Potthast, M. Hagen, M. Völske, and B. Stein. Crowdsourcing interaction logs to understand text reuse from the web. *Proc. of ACL 2013*.

[17] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. *Proc. of CIKM 2004*.

[18] R. White, G. Muresan, and G. Marchionini, editors. *Proc. of SIGIR workshop EESS 2006*.

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