A Corpus of Realistic Known-Item Topics With Associated Web Pages in the ClueWeb09

Matthias Hagen, Daniel Wägner, and Benno Stein

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

Abstract. Known-item finding is the task of finding a previously seen item. Such items may range from visited websites to received emails but also read books or seen movies. Most of the research done on known-item finding focuses on web or email retrieval and is done on proprietary corpora not publically available. Public corpora usually are rather artificial as they contain automatically generated known-item queries or queries formulated by humans actually seeing the known-item.
In this paper, we study original known-item information needs mined from questions at the popular Yahoo! Answers Q&A service. By carefully sampling only questions with a related known-item web page in the ClueWeb09 corpus, we provide an environment for repeatable realistic studies of known-item information needs and how a retrieval system could react. In particular, our own study sheds some first light on false memories within the known-item questions articulated by the users. Our main finding shows that false memories often relate to mixed up names. This indicates that search engines not retrieving any result on a known-item query could try to avoid returning a zero-result list by ignoring or replacing names in respective query situations.
Our publically available corpus of 2,755 known-item questions mapped to web pages in the ClueWeb09 includes 240 questions with annotated and corrected false memories.

1 Introduction

In the field of information retrieval, *known-item search* is the common task of re-finding a previously accessed item. Types of known items include visited web sites, received or written emails, stored personal documents, but also read books, seen movies, or songs heard on the radio.

In contrast to informational or transactional searches, which can have a multitude of viable results, the goal of a known-item search usually is to retrieve a single, specific item (or syntactic/semantic aliases of it) [6]. In some cases a hub that is "one step away from the target [item]" can also be a less desirable, but still acceptable result [6]. An example for such a hub could be a web page clearly linking to the page a user is looking for or the track listing of a music album, with one of the songs being the desired known item.

Consequently, the number of relevant or useful results tends to be much smaller for known-item queries than for other query types. On the other hand, the

user often has a larger amount of information which can be used to narrow down the results of a known-item query. These two points, the number of acceptable results and the available knowledge, are two main factors that separate known-item searches from other search tasks.

While a large amount of available information can make it easier to re-find a known item, particular attention needs to be paid to incomplete or false memories. Studies have shown that humans remember some kinds of details better than others [4, 11, 16]. For example, a user looking for a movie might misremember details about the setting (by thinking that it took place in Ireland, rather than Scotland), the cast (by confusing Danny Glover with Morgan Freeman) or misquote a specific line (Darth Vader never says the exact phrase "Luke, I am your father" in "The Empire Strikes Back"). False memories are problematic in that they can lead to the desired item being excluded from the results of a formulated query containing the false memory. A search engine taking the query as is (i.e., including the false memory) might not find any matching result. Presenting an empty result list should be avoided since they harm user experience. Thus, taking care of false memories on search engine side helps to avoid such situations (e.g., the search engine could try to correct the false memory or remove it from the query in a "did you mean"-way [13]). Our study will focus on identifying and characterizing typical false memories. One of our main results shows that searchers often mix up person names when looking for movies or songs.

Current research on the topic of known-item retrieval relies heavily on corpora of known-item queries and their respective known items. Unfortunately, many of those corpora (1) are proprietary and not publicly available, (2) consist of automatically generated queries [2, 17, 10], or (3) consist of queries generated manually from a known item itself, in a human computation game [18].

Hauff et al. [14] characterized proprietary corpora as problematic since they do not allow for repeatable experiments. Hauff et al. also stated that queries generated from the known item itself, whether automatically or manually, are rather artificial and not representative of real-world user queries since they make unrealistic assumptions: randomly failing memory in automatic query generation or almost perfect memory in human computation games where the known item actually is displayed during or shortly before query formulation. To provide an alternative to these existing corpora, Hauff et al. proposed the creation of a known-item topic set built from questions posted by users of the Yahoo! Answers platform,[1] with the aim to address the lack of public data and the unrealistic approaches to query generation they identified in prior work [14]. As a proof of concept, 103 questions by Yahoo! Answers users were crawled. Among those, 64 information needs were manually assessed, consisting of 32 website and 32 movie known items. Interestingly, even a handful of false memories could be identified.

In the paper at hand, we significantly expand on the ideas of Hauff et al. and build a large-scale corpus with a wider coverage of different information needs, suitable for use in further research. Studying known-item information needs from Yahoo! Answers, we analyze false memories in realistic situations. To

[1] `http://answers.yahoo.com`

ensure the usability of our experiments in a broader context, we only examine known items with a related web page in the ClueWeb09 corpus. For non-website items, like movies or books, this is usually their corresponding entry in the English Wikipedia. The corpus consisting of 2,755 known-item questions mapped to web pages in the ClueWeb09 corpus (including 240 questions with annotated and corrected false memories) is publically available.[2]

The paper is organized as follows. Section 2 describes related work on known-item finding. We present our methodology of corpus construction in more detail in Section 3. First analysis results are reported in Section 4, followed by conclusions and ideas for future work in Section 5.

2 Related Work

We first describe studies that investigated the process of re-finding in different contexts and then focus on studies of known-item querying in particular.

Re-finding Behavior Blanc-Brude and Scapin [4] conducted a study investigating the ability to recall attributes of a user's own documents (both paper and digital ones) and whether the users could re-find those documents in their work place. The documents were classified as *old* (last access six or more months ago), *recent* (last access within the last six months) and *recurrent* (regularly accessed). The findings show that the study participants were most often mixing true and false memories when being asked to recall the title and keywords of a document in question. For 32% of the documents the recalled keywords were correct, while for 68% they were only partially correct. Recalling the title was even more difficult: 33% correctly recalled document titles versus 47% partially correct and 20% completely false recollections. Location, format, time, keywords, and associated events were remembered most frequently; still, many of these attributes, particularly keywords, time, and location were often only partially remembered or the recollections were incorrect.

Elsweiler et al. [8, 9] performed user studies to investigate what users remember about their email messages and how they re-find them. The most frequently remembered attributes of emails were the topic, the reason for sending the email, the sender of the email and other temporal information. In the evaluation, no indication was given if the memories were (partially) false or correct but another finding, in line with research in psychology, was that memory recall declines over time. Emails that had not been accessed for a long time were less likely to have attributes remembered than recently read emails. That users are indeed accessing old documents was shown by Dumais et al. [7]: up to eight years old documents were sought by users in a work environment.

In case of re-finding behavior on the web, people also often do re-find and revisit pages they have accessed a couple of days ago [1]. The last visited documents of a previous session are typically pages to be re-found at the beginning of a later session and people tend to formulate better (i.e., shorter) queries over time, when they access the same item several times [20].

[2] `http://www.webis.de/research/corpora`

A range of studies [3, 5] showed that users in general prefer to browse and to visually inspect items in order to re-find a target document instead of relying on provided text-based search tools. It is then argued that the current personal information management search tools are not sophisticated enough to deal with what and how users remember aspects of the target documents. This is probably also true for the web where the typical interface for re-finding also is a simple keyword-based search box—that still is highly effective for many tasks.

In our scenario, we also consider known-items that have a corresponding web document but we will mostly focus on known-items that have been seen more than just a couple of days ago. We study known-item information needs submitted to a popular question answering platform. Similar to most of the cited studies, also in our study users face the problem of false memories and problems in articulating their need as a query or question when the item was accessed a longer time ago. In contrast to many other search related studies, our corpus of 2,755 known-item information needs connected to ClueWeb09 documents is publically available in order to support further research.

Known-Item Query Generation Since no large-scale query logs with known-item queries are available, different approaches to generate known-item queries have been proposed ranging from automatic generation to human computation games. For instance, the automatic known-item topic generation approach by Azzopardi et al. [2] works as follows: a known-item / query pair is generated by first selecting a document from the corpus in the role of the known item and by then deriving a corresponding query. The query terms are drawn from the selected document according to particular probability distributions (e.g., the most discriminative terms are selected with a higher probability) while adding some random noise models memory problems. This process was also adapted for the case of personal information management and emails [17, 10]. Since such documents usually consist of different fields—emails for instance have a sender, a title, a sending date and a body—, the query terms are drawn from the fields with different probabilities to mimic human memory.

Rather than using automatic query generation, Kim and Croft [18] employ a human computation game to create more "natural" queries. Study participants were shown the known item in question and shortly thereafter they were asked to create a query that retrieves the known item as high as possible in the ranking of a standard retrieval engine. However, even though showing the known item to a user may entail natural queries (i.e., queries created by humans), it does not fully include the concept of false memories.

Hauff et al. [15, 14] emphasize the importance of realistic query generation scenarios including false memories when studying search behavior in the known-item setting. They conclude that none of the existing query generation approaches are really realistic as the studied corpora are either proprietary and not publicly available, or consist of automatically generated queries, or consist of queries generated manually from a known item itself. Following Hauff et al.'s suggestions [14] our proposed methodology addresses these problems: we collect a set of 2,755 known-item topics from a popular question answering platform.

The known-item topics are based on real information needs by users having problems remembering the known item fully or correctly. Our first results will show what the main issues are with false memories in these cases.

3 Corpus Construction

As discussed in the related work section, the existing approaches to constructing publically available known-item corpora tend to yield rather artificial results. We propose our new Webis Known-Item Question Corpus 2013 (Webis-KIQC-13) as an alternative to those corpora, with the goal of providing a freely available known-item corpus based on real information needs expressed by real humans and with linked items in the popular ClueWeb09 corpus. In principle, our corpus construction follows the suggestions of Hauff et al. [14]. We select questions and answers from a question answering platform where the desired known-item has a corresponding web page in the ClueWeb09 corpus. For the sampled questions and answers a manual annotation identifies the known-item intent and whether a false memory is contained (with manually annotated corrections). This section provides the details on the process of corpus construction.

3.1 Crawling Known-Item Topics from Yahoo! Answers

Web-based community question-answering (cQA) services allow users to pose questions to other users, rate answers by others and receive rewards for providing good answers to open questions. We chose the Yahoo! Answers platform for our purpose of retrieving known-item topics since it provides a public API and a broad range of information needs submitted by many different users. Users are able to submit questions expressed in natural language. These are then opened for other users to propose answers or vote for the best answer to a question. If no best answer gets selected by the asker during the open period, the community votes given by other users potentially determine the chosen answer. In both cases, the question is marked as *resolved*. If no best answer can be chosen through either method, the question is labeled as *undecided*.

For building our known-item topic set, we use the public Yahoo! Answers API, which for example allows retrieving up to 1,050 question entries matching a given keyword query. Our primary focus is on retrieving questions on three types of known items that are often searched for: websites, movies, and musical works (songs and music albums). Nine separate API queries were formulated for each of the three types; to provide a broader range of topics, ten additional queries for other types of known-item information needs were formulated, such as re-finding a book or TV series. Examples of the used API queries are shown in Table 1. To avoid the effect of low quality answers, we only sampled resolved questions from the Yahoo! Answers API. On January 21, 2013, the 37 distinct search queries were submitted to the Yahoo! Answers API, which resulted in a combined set of 24,765 unique questions.

In a second step, the comments and information about who voted for a best answer (community or asker) were scraped from each question's HTML version

Table 1. Examples of search queries used to retrieve from Yahoo! Answers.

`(remember) AND (title) AND (movie)`
`(forgot) AND (name) AND (film)`
`(forgot) AND (title) AND (song)`
`(forgot) AND (url) AND (website OR (web site))`
`(remember OR forgot) AND (name OR title) AND (book)`

on the Yahoo! Answers website since they were not contained in the API results. The comments that the asker added to an answer can sometimes be a valuable indication of whether an answer actually contained the searched item and best answers selected by the original asker are a better indication of a correctly found known item than are community votes. Note that also the Yahoo! Answers point system promotes that the asker should select a best answer if there is one. In this case, 3 points are gained while a community vote (that is likely when the desired item is in an answer) does not yield any points. Six questions returned by the API were no longer accessible; among the remaining 24,759 questions only 8,825 questions had their best answer chosen by the original asker. These were kept for manual assessment.

3.2 Assessment of the Crawled Questions

The crawled questions and answers were manually assessed to ensure that they represent satisfied known-item information needs and that they correspond to some website in the ClueWeb09 corpus. The assessors were presented with a form that contains the data fields retrieved by the API query and HTML scraper and additional fields that are to be filled out manually. An external window provides a web view, which allows the assessor to view questions as they are presented to Yahoo! Answers users, to follow hyperlinks and to perform web searches. We had two assessors who checked each of the crawled questions independently. The assessors discussed their decisions afterwards for the few questions where they did not agree initially to reach a consensus.

Assessment of Question Intent For each of the 8,825 questions with a best answer chosen by the asker, it was first judged whether the intent was to re-find a previously known item, and whether the answer was the desired known item.

For example, questions like "What is the weirdest movie you remember from your childhood?" or "What songs are similar to 'Remember The Name' by Fort Minor?" match our API queries but are posed to initiate a discussion or to receive a recommendation, rather than to satisfy a known-item information need.

For some known-item questions, the asker commented that an answer did not contain the known item, but still chose it as the best answer. This would happen if the answer was still useful to the asker (e.g., recommending a similar item), or merely so the asker would gain some points. In both cases, the questions are omitted from our corpus, as the desired known item could not be determined.

In total, 5,419 questions were discarded in this step, further narrowing down the topic set to 3,406 known-item information needs. Although similar search terms were chosen for all types of items, the proportion of discarded questions varied widely. While only about 35% of movie questions had to be discarded, for websites it were more than 95%. Possible explanations are the following.

- The default behavior of the API, to search in both the question and the answer, led to a large number of unwanted results. For instance, one of the website API queries returned almost one-hundred site support questions answered by the same user, with the same or similar stock answers containing every part of the search term. All of these had to be discarded.
- Askers may be less interested in re-finding a specific website than they are for other item types. Frequently, users are also content with an alternative website offering the same functionality, even if it is not the known item.
- The search terms in our API queries may be ill-suited for finding known-item website questions. The analysis of other cue phrases could be an interesting path for investigation in future research.
- Website re-finding questions in general may be less often submitted to Yahoo! Answers, compared to those for movies, music, or books.

Website re-finding information needs were originally supposed to form a major part of our Webis-KIQC-13. However, only 82 out of 1,706 website known-item questions remain after the intent assessment step.

Mapping of Known Items to their ClueWeb09 ID In the next step, the assessors checked whether a known item's URL is included in the ClueWeb09. For website questions, this would be the website's URL itself. For most other types of items, we decided that an appropriate URL would be the corresponding article in the English Wikipedia, if there is one. It should be noted that a known item may have multiple semantically or syntactically equivalent aliases [6]. For example, a movie can have both a Wikipedia article and a corresponding IMDb entry, or a notable website may in turn have a Wikipedia article. In these cases, the more appropriate known-item URL in the ClueWeb09 was preferred (e.g., the URL containing more content on the known item). Also, as noted by Broder et al. [6], a so-called *hub*-type result, which is one step away from the target, can be an acceptable, although less desirable result. Examples where hub-type results were deemed acceptable by our assessors include songs not represented through a Wikipedia article of their own, but through the album they were released on, or specific pages on a website where only the main page is in the ClueWeb09.

We used the publically available ChatNoir API [19] that easily maps an item's URL to the corresponding ClueWeb09 ID. Still, the mapping of URLs to ClueWeb09 IDs often had to be done manually by the assessors as a movie or song title often could not directly be translated to a Wikipedia-URL and also the decision of whether a hub-result is contained in the ClueWeb09 had to be determined manually. For 651 out of the 3406 known items, no ClueWeb09 entry could be identified; only the 2,755 known-item questions with matching ClueWeb09 entries form our Webis-KIQC-13. Most of the discarded questions were posed for

Table 2. Examples of tagged false memories in Yahoo! Answers questions.

Known item	False memory / Correction
Shooter (film)	[...] Morgan freeman offers him a job to kill a person [...] wrong actor: Danny Glover, not Morgan Freeman
Tokio Hotel	What's the english emo rock band [...] They are american [...] origin: German band, not English or American
An American Tail	[...] a Disney cartoon about a little mouse [...] company: Amblin Entertainment, not Disney
theforgottenlair.net	[...] it went somethin like the underground lair [...] URL: "forgotten", not "underground"

known items more recent than the 2009 crawl date of the ClueWeb09. Given the age of the ClueWeb09 corpus, we expected such an outcome. The differences in coverage over time will be further analyzed in Section 4.

Annotation of False Memories Finally, the assessors determined whether a known-item question contained false memories. In these cases, the assessors tagged the question as such and added a short annotation documenting the type of error, a correction, and the misremembered property. Some examples of false memories in Yahoo! Answers questions and their annotated corrections are shown in Table 2. Of the 2,755 known-item questions in the Webis-KIQC-13 corpus, 240 (8.7%) contain at least one false memory.

Summary Although we started from a base of 24,759 unique questions retrieved from the Yahoo! Answers API, the final topic set consists of only 2,755 suitable known-item information needs (11.1% of the original crawl). This is mostly due to the decision to exclude questions decided by community vote, which account for about two in three questions across all crawled categories. A summary of the items removed in the assessment steps is given in Table 3. The large amount of non-known-item questions that we had to discard for some topics is a little surprising. Possible explanations for the case of website information needs have already been hypothesized above. These explanations might, to a lesser degree, be applicable to other categories as well.

The amount of false memory effects identified in the corpus met our initial expectations to be in the range of 5–10% that was also found in the small-scale study by Hauff et al. [14]. The actual number of false memories may be even higher. As the annotators mostly had to rely on the answer text and the known item's corresponding ClueWeb09 document, it is likely that they missed false memories that were not explicitly mentioned therein.

As argued by Azzopardi et al. [2], the manual construction of a known-item corpus on the scale of our Webis-KIQC-13 is a laborious and time-consuming process. Our two assessors together spent approximately 400 hours on the evaluation of the 8,825 questions that had an answer chosen by the asker which translates to an average of about 80 seconds per question.

Table 3. Summary of the removed/remaining items during assessment. Note that the column "Total" also includes additional categories like books etc.

	Movies	Music	Websites	Total
Retrieved questions	5,896	6,481	5,343	24,759
Best answer chosen by voters	-3,718	-4,112	-3,637	-15,934
Best answer chosen by asker	2,178	2,369	1,706	8,825
Not known-item questions	-768	-1,451	-1,624	-5,419
Known-item questions	1,410	918	82	3,406
Not in ClueWeb09	-250	-219	-20	-651
In ClueWeb09	1,160	699	62	2,755
Containing false memories	81	74	4	240

4 Corpus Analysis

We provide a first analysis of the known-item information needs contained in our Webis Known-Item Question Corpus 2013 (Webis-KIQC-13) and their associated properties. We briefly analyze the coverage of the ClueWeb09 corpus and then focus on the types of false memories exhibited. These false memory analyses and the release of our corpus are meant as an enabler for research on the influence of false memories on retrieval processes. By no means, our first analyses can be conclusive but will shed some light on very interesting directions for future work.

4.1 ClueWeb09 Coverage

The ClueWeb09 has been crawled from the live web in January and February 2009. We examine the coverage of the known-item questions by the time of their submission to Yahoo! Answers. Note that, although the newer corpus ClueWeb12 is much younger with a crawling period between February 10, 2012 and March 10, 2012, unfortunately it does not contain Wikipedia and thus lacks the main source of known-item URLs we are aiming for.

The left part of Table 4 presents the relative ClueWeb09 coverage of the retrieved known item queries per year. The steep increase in the number of retrieved known item questions in 2008 can probably be related to an increase in Yahoo! Answers usage. Beginning from 2009, the ClueWeb09 coverage predictably decreases due to the occurrence of known items that did not exist at the time of the ClueWeb09 crawl (e.g., newer movies). While in 2007 a record high of 92.2% could be achieved, the known-item coverage fell to only 71.9% for 2012. By a closer analysis of the known-item questions, we noticed that there were two major groups of re-finding needs that are influenced differently by the ClueWeb09 crawling date. We have (1) questions for items that have not been accessed for a long time (e.g., users searching for the favorite movie of their childhood), and (2) questions for items that have only been incompletely accessed more recently (e.g., by watching the trailer of a movie the other day). Obviously, the web corpus crawling data has a much higher impact on the latter type.

Table 4. ClueWeb09 coverage of the originally crawled 3,406 known-item questions by year and domain type.

	2006	2007	2008	2009	2010	2011	2012	Wikipedia	IMDb	Others	No link
Webis-KIQC-13	68	176	369	701	578	477	364	2,618	3	134	-
Not in ClueWeb09	8	15	60	112	148	140	142	405	66	94	86
Total	76	191	429	813	726	617	506	3,023	69	228	86
Coverage	89.5%	92.2%	86.0%	86.2%	79.6%	77.3%	71.9%	86.6%	4.3%	58.8%	0%

Further, we also examine the domains of the ClueWeb09 documents used to represent the known items. The right part of Table 4 shows the frequency with which websites were chosen by the assessors. As can be seen, Wikipedia is the first source the assessors checked when searching for a known item's URL, and the majority of known items were matched to their article there. This decision was made since the ClueWeb09 corpus contains a nearly complete dump of the English Wikipedia at the time of its crawl. At the time of our assessment, 3023 known items either had a Wikipedia article of their own or, as per Broder et al.'s definition [6], a *hub*-type result on the live web. However, for 405 out of them, the Wikipedia article is not part of the ClueWeb09. These 405 were then checked against IMDb or other domains. However, only three out of 69 IMDb entries found on the live web were actually part of the ClueWeb09. Note that in 86 cases, the assessors could not even find a suitable document representing the known item on the live web. These were usually items like poems or songs not released on some album with a Wikipedia entry.

4.2 False Memories

At least 240 of the 2,755 known items in the Webis-KIQC-13 contain some kind of false memory. Categories of false memories were defined ad-hoc by the assessors and were unified in a second pass over the information needs with false memories. Given the search terms used to retrieve our topic set, most of the information needs relate to works of art and entertainment. The most common types of memory errors are shown in Table 5, with an explanation and their number of occurrences. Note that especially the categories relating to persons (character, artist, and actor) with their total amount of 67 false memories form the biggest problem users had in articulating their information need. These categories mostly relate to movie and music questions. Especially for music questions, the lyrics category is another big source of problems. Some text might be mixed up or only remembered in a misheard form and thus can not lead to a good retrieval result.

Our first, and still very basic, analyses reveal two important findings for retrieval systems when taking false memories into account. First, when a query or question including person names does not yield any search result, it is not unlikely that the name is a false memory. A retrieval system could then support the user by leaving out the name for retrieval or suggesting related names (e.g., other actors) that would yield results. Second, queries or questions including lyrics tend to contain false memories. Incorporating sophisticated phonetic

Table 5. Common types of false memories in the Webis-KIQC-13.

Category	False memories relating to ...	#
character	attributes of character in a work of fiction	34
lyrics	lyrics of song or poem	29
title	title of work	27
format	way work was released	21
artist	wrong attribution of artist to musical work	22
time	time a work has been produced or released	18
origin	geographical background of a work or artist	15
actor	wrong attribution of actor to movie or series	11
plot	key elements of a work's plot	9
setting	time or place a work is set in	9
company	company involved in production of item	6
scene	single scene in movie or series	5
prop	object in movie or theater play	5
mix-up	confusing attributes of two items	5
URL	URL of website	4

similarities at retrieval system side might be a research direction to support the frequent case of false memories in form of misheard lyrics (e.g., "Stayin' Alive" by the Bee Gees is often misheard as "Steak and a Knife").

5 Conclusions

Our Webis Known-Item Question Corpus 2013 (Webis-KIQC-13) enables a new approach to the evaluation of known-item retrieval tasks, based on using real information needs with a clearly stated intent of known-item re-finding. We believe that by constraining the topic set to answers selected as correct by their asker, we could minimize the error in our known-item mappings. In connection with the ClueWeb09 corpus, this topic set allows for repeatable and realistic testing of known-item information needs. The corpus is freely available.[3]

One direction we envision as particularly promising besides general known-item question analyses is the false memories we annotate in the corpus. They often relate to important details of the known item being sought. The investigation of these false memories is an interesting path for future research. Based on the false memories contained in queries, search engines might not find any reasonable result. To avoid such zero-result lists, the false memories could be identified by to-be-developed techniques and then replaced or removed in a did-you-mean manner [13].

The annotated false memories could also be used to examine the recall of different kinds of information in audiovisual media since most of the search terms we used to crawl questions from the Yahoo! Answers API acquired known-items from the categories Arts & Humanities as well as Entertainment & Music. This

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

places a large number of information needs in our Webis-KIQC-13 close to the field of media or video retrieval, although from a different vantage point.

Incorporating other types of known items that users might search for, such as geographical landmarks or electronic devices, is an interesting direction for future corpus enrichment to provide a representative sample of all potential known-item intents. Especially interesting in that respect would be the inclusion of many more website items. For that category, our search terms that yielded acceptable results on other categories hardly returned usable known-item information needs.

Although our corpus was originally developed as a testbed for known-item search tasks, other uses could be considered as well. Since our Webis-KIQC-13 is publically available and is linked to the widely-used ClueWeb09 corpus, repeatable research on web requests in the known-item domain is possible.

References

1. E. Adar, J. Teevan, and S. T. Dumais. Large scale analysis of web revisitation patterns. In *CHI 2008*, pp. 1197–1206.
2. L. Azzopardi, M. de Rijke, and K. Balog. Building simulated queries for known-item topics: an analysis using six european languages. In *SIGIR 2007*, pp. 455–462.
3. D. Barreau and B. Nardi. Finding and reminding: file organization from the desktop. *ACM SIGCHI Bulletin*, 27(3):39–43, 1995.
4. T. Blanc-Brude and D. L. Scapin. What do people recall about their documents?: Implications for desktop search tools. In *IUI 2007*.
5. R. Boardman and M. Sasse. Stuff goes into the computer and doesn't come out: a cross-tool study of personal information management. In *CHI 2004*, pp. 583–590.
6. A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.
7. S. T. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff I've seen: A system for personal information retrieval and re-use. In *SIGIR 2003*, pp. 72–79.
8. D. Elsweiler, M. Baillie, and I. Ruthven. Exploring memory in email refinding. *ACM Trans. Inf. Syst.*, 26(4):1–36, 2008.
9. D. Elsweiler, M. Baillie, and I. Ruthven. What makes re-finding information difficult? A study of email re-finding. In *ECIR 2011*, pp. 568–579.
10. D. Elsweiler, D. E. Losada, J. C. Toucedo, and R. T. Fernández. Seeding simulated queries with user-study data forpersonal search evaluation. In *SIGIR 2011*, pp. 25–34.
11. D. Elsweiler, I. Ruthven, and C. Jones. Towards memory supporting personal information management tools. *JASIST*, 58(7):924–946, 2007.
12. R. Gunning. *The technique of clear writing*. McGraw-Hill, 1952.
13. M. Hagen and B. Stein. Applying the user-over-ranking hypothesis to query formulation. In *ICTIR 2011*, pp. 225–237.
14. C. Hauff, M. Hagen, A. Beyer, and B. Stein. Towards realistic known-item topics for the ClueWeb. In *IIiX 2012*, pp. 274–277.
15. C. Hauff and G.-J. Houben. Cognitive processes in query generation. In *ICTIR 2011*, pp. 176–187.
16. L. Kelly, Y. Chen, M. Fuller, and G. J. F. Jones. A study of remembered context for information access from personal digital archives. In *IIiX 2008*, pp. 44–50.
17. J. Kim and W. B. Croft. Retrieval experiments using pseudo-desktop collections. In *CIKM 2009*, pp. 1297–1306.
18. J. Kim and W. B. Croft. Ranking using multiple document types in desktop search. In *SIGIR 2010*, pp. 50–57.
19. M. Potthast, M. Hagen, B. Stein, J. Graßegger, M. Michel, M. Tippmann, and C. Welsch. ChatNoir: A search engine for the ClueWeb09 corpus. In *SIGIR 2012*, p. 1004.
20. S. K. Tyler and J. Teevan. Large scale query log analysis of re-finding. In *WSDM 2010*, pp 191–200.