Towards realistic known-item topics for the ClueWeb

Claudia Hauff*
Delft University of Technology
Delft, The Netherlands
c.hauff@tudelft.nl

Matthias Hagen Anna Beyer Benno Stein
Bauhaus-Universität Weimar
99423 Weimar, Germany
<first name>.<last name>@uni-weimar.de

ABSTRACT

Known-item finding is the task of re-finding and re-accessing an item previously seen. Typical examples of known items include accessed Web sites, received emails, or documents on one's personal desktop. Current research on known-item finding heavily relies on corpora of known-item queries and the respective known items. However, many existing corpora are proprietary and not available to the public (in particular those derived from Web query logs), a fact which does not allow for repeatable research. The existing publicly available corpora either contain automatically generated queries or queries that were manually generated while seeing the known item itself. Hence, we consider these public corpora to be rather artificial in nature.

In this paper, we propose a methodology to create a known-item topic set that is much more realistic and that is built on top of a large-scale public test corpus. From know-item questions posted on the popular Yahoo! Answers platform we extract queries for known-items in a crowdsourcing setup. Since we ensure that all the known-items correspond to Web pages in the publicly available ClueWeb09 corpus (a large static Web crawl), we provide an environment for repeatable realistic Web-scale known-item searches.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval
General Terms: Experimentation
Keywords: ClueWeb, known-item

1. INTRODUCTION

A vital component of research in information retrieval is the testing of research ideas on realistic test collections. Creating such test collections is both time-consuming and cost-intensive. For this reason, several initiatives, such as TREC,[1] CLEF,[2] and NTCIR[3] have been set up over the years. They provide researchers with standardized test corpora and retrieval tasks. In text retrieval the most prominent retrieval task is the ad-hoc (Web) retrieval task, but other tasks such as named entity retrieval, passage retrieval and known-item retrieval have also been investigated. The latter task is the one we concern ourselves with in this paper.

Since the introduction of the Web more than two decades ago, we have seen a massive and unprecedented digitization of our lives. We increasingly rely on digital communication channels, such as email or social networks, to keep in touch, to work, and to handle every-day tasks. With rapidly decreasing storage costs, larger amounts of our digitally distributed information are being stored including bank statements, project reports, e-books, videos and bookmarks among others. Moreover, information access is becoming more heterogeneous as we read emails on our smartphones and laptops, we write reports on our tablet computers and PCs, etc. Today, most people will have data in their personal digital archives (PDA) that spans less than a decade. However, if we look ahead, in another ten years, PDAs will have grown by several magnitudes. Re-finding an item that the user has accessed before, either in his PDA or on the Web, is a process known as *known-item* retrieval. It is becoming an increasingly important task due to the just described increasing amount and increasing longevity of our data.

Research in known-item retrieval is conducted along two lines: naturalistic studies of re-finding behavior (on the Web or in PDAs) and studies involving an artificial component to create a topic set. We follow the second approach here, but alleviate the issues of previous approaches by utilizing real information needs.

Studies investigating the re-finding behavior on the Web commonly exploit large-scale query logs [14]. A major drawback of such studies though is the proprietary nature of the log data, which does not allow further investigation by other researchers. Research in search technologies related to personal information management (PIM) is similarly hampered by the lack of public test corpora. To alleviate this problem, automatic [1, 11] and human computation game [12] based topic set generation approaches have been proposed in the past. Given a test corpus that resembles a generic user's personal or work Desktop, a document of the test corpus is selected as the "known item" for which a query is created. The automatic approaches construct queries by selecting terms

*This research has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no ICT 257831 (ImREAL project).

IIiX 2012, Nijmegen, The Netherlands

[1]Text REtrieval Conference
[2]Cross Language Evaluation Forum
[3]NII Test Collection for IR Systems

of the document in question according to particular rules. In the human computation game scenario, the document in question is shown to human study participants who create queries with the goal to get the item ranked as high as possible. Note that the human participants get the interesting document shown, they do not need to remember it.

The two outlined known-item topic creation approaches assume either (i) a perfect human memory where users remember the document's content fully and correctly and it is only a matter of selecting the "right" keywords to create a good query (in the human computation game approach), or, (ii) a human memory that fails randomly (in the automatic query generation approach). Human memory is neither perfect nor failing randomly, however. On the contrary, forgetting details of a document's content over time or incorrectly remembering attributes of a document are pretty common. Indeed, research in so-called *false memories* is an important field of study in psychology, which is often motivated by the question of eyewitness reliability [13] and the correct recall of childhood experiences [10].

In this paper, we argue that for known-item retrieval to be more realistic, topic generation approaches should consider the imperfection of human memory and the tendency to create false memories. This argument is also supported by user studies in PIM, which have shown that users recall different aspects of their stored documents to different degrees [8]. This is of importance as developing new search algorithms based on perfect memory queries or randomly failing memories may lead to false estimates of the algorithms' abilities. For instance, the TREC Enterprise track 2005 [6] contained a known-item task where the best systems retrieved the known item within the top ten ranks for more than 80% of all queries, which implies very well-performing known-item retrieval algorithms. Some of the known items in question, though, where ten year old emails (at the time of topic creation), which are unlikely to be remembered correctly in a realistic search setting [9].

We introduce a topic generation method that addresses the issues we have identified, namely a lack of public data sets and unrealistic query generation approaches that do not take false memories into account. Our methodology is as follows: we utilize the most recent Web TREC corpus (ClueWeb [5]) and identify information needs posted by real users that attempt to re-find two types of items: movies and Web sites. We extract these information needs from Yahoo! Answers[4] and include only those where the wanted item is part of the ClueWeb corpus (in case of the movies we focus on the respective Wikipedia or IMDB Web site). In effect, we thus create a new public topic set that can be used with the ClueWeb corpus, focusing on information needs where users have problems re-finding previously discovered items.

The rest of the paper is organized as follows: Section 2 describes related work in PIM and in known-item query generation approaches. We present our methodology in more detail in Section 3. First experimental results are reported in Section 4, followed by the conclusions in Section 5.

2. RELATED WORK

We first present a number of studies that investigated the influence of false memories in PIM and then describe the existing query generation approaches in greater detail.

[4]http://answers.yahoo.com/

False memories in PIM.

Blanc et al. [3] describe the results of a user study, in which the ability to recall attributes of the users' own documents (both paper and digital ones) and their ability to re-find those documents in their work place was investigated. It was observed that the study participants were most often mixing true and false memories when being asked to recall the title and keywords of the 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 more difficult: 33% correctly recalled document titles versus 47% partially correct and 20% completely false recollections.

Elsweiler et al. [8] performed a user study to investigate what users remember about their email messages and how they re-find them. The most frequently remembered attributes of emails were found to be the topic, the reason for sending the email, the sender of the email and other temporal information. 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 on their Desktop has been shown in [7], where up to eight years old documents were sought by users in a work environment.

In general it has been found across a range of studies (e.g. [2, 4]) that in PIM re-finding, users prefer to browse and to visually inspect the target folder in order to find the target document instead of relying on the provided Desktop search tools. It is argued that the current PIM search tools are not sophisticated enough to deal with what and how users remember aspects of the target documents.

Query generation approaches.

The automatic known-item topic generation approach originally proposed by Azzopardi et al. [1] 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 the 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). In [11], this process was adapted for the particular case of PIM, where documents usually consist of different fields—emails for instance have a sender, a title, a sending date and a body. Essentially, the query terms were drawn from certain fields with a higher probability than from others. In [1], lapses in memory are modelled with a random noise component. This is not a very realistic setup as in reality false memories do not occur completely at random.

Rather than generating the queries automatically, Kim et al. [11] also presented a human computation game where study participants create queries for the selected known item documents manually: the participants were shown the known item in question and they were asked to create a query that retrieves the known item with a standard retrieval engine as high as possible in the ranking. Such a setup, where the known item is shown to the user while she creates the query, may entail natural queries (i.e., queries created by humans) but does not include the concept of false memories.

Our proposed methodology addresses these problems: we develop a set of topics that are manually created and based on real information needs posed by users having problems remembering the known item fully or correctly.

"cannot remember" website	forgot url
"cannot remember" movie	forgot film title
"cannot find" website	"help me find" website

Table 1: Examples of cue phrases that the three assessors used to search for suitable information needs on Yahoo! Answers.

3. METHODOLOGY

Both automatic query generation and human computation games yield rather artificial known-item queries, as argued in the previous section. By contrast, our new corpus shall be based on real information needs of real humans. As a source of such information needs we choose the Yahoo! Answers question answering platform on which humans can express a question in natural language, while other users can answer these questions. Answers can be voted as a best answer to the respective question, which is often done by the questioner herself and which labels the question as "resolved." As we want to build a corpus of queries with respective known results, we focus on resolved questions only.

To build the topic set we have used the Yahoo! Answers API. We focus on two types of known items that are often searched for: Web sites and movie titles. In both instances, a user might still remember some facts (e.g., the basic movie plot or parts of the Web site's URL) but cannot remember the desired item itself. Three assessors manually searched for suitable questions by way of cue phrases. Examples of the cue phrases are shown in Table 1.

Within a manual assessment step, we then judge whether a crawled query's intent is known-item re-finding or something else. For instance, the question "Forgot my password, where is the URL to reset it?" is obviously not a known-item question in our notion. Furthermore, we also check whether the known item's URL (in case of movies, this is the respective Wikipedia or IMDB page) is part of the ClueWeb09 corpus. Information needs for known items that are not covered by ClueWeb are discarded.

The obtained known-item re-finding questions are still written in natural language and are rather lengthy: the shortest found information need contains 17 words, the longest one contains 194 words. Hence, the problem remains to obtain realistic Web queries from the crawled questions. Therefore, we set up a crowdsourcing task, where we ask humans to formulate a Web query from the natural language questions. In the task description, neither the known item nor the answer to the Yahoo! Answers question are contained, but the workers are informed that they should formulate a Web query from the given information need assuming they forgot the Web site or the movie title.

Our resulting topic set contains the queries generated via crowdsourcing along with the known items searched for. The requirement of the known item being contained in the ClueWeb09 corpus enables our known-item topic set to be used in repeatable experiments as it will be made publicly available. It should be emphasized that in this way, we generate a "difficult" known-item query set: users generally do not post their information needs on Yahoo! Answers if they could easily retrieve the item themselves with the help of a search engine. Often, the imperfections of human memory play a role, rendering such queries more realistic since they specifically include false memory aspects.

4. EXPERIMENTAL RESULTS

Three assessors (experienced search experts) searched the Yahoo! Answers portal for suitable information needs. To demonstrate the feasibility of our corpus construction approach, the assessors identified 64 information needs out of 103 crawled Yahoo! Answers queries, with 32 being Web site known items and 32 being movie known items. The main reason that invalidates Web sites as known items is the fact that they are not contained in the ClueWeb corpus; however, this applies less for movie questions since ClueWeb contains the respective Wikipedia entries for most movies released prior to 2009.

The information needs crawled from Yahoo! Answers are often elaborate descriptions of the wanted item, for example: *"I am looking for the website that takes your photo, and puts it on a billboard of another photo, or a gallery, or a magazine cover, does anyone know the website that does this?"*

Since essential information about the desired item is missing, the information needs tend to be verbose. While the shown example is not spoiled by false memories, examples of information needs that include a false memory component are given in Table 2.

Once the information needs were identified, we recruited workers for the crowdsourcing setup. Altogether 11 subjects participated in our study, processing between 1 and 64 (all) of the presented information needs. Each information need was processed by at least two subjects and thus, at least two queries were generated for each one. On average, the created queries contain seven words, which render them longer than the average Web query—but still considerably shorter than the original information need. Furthermore, the variance in the length is high since some queries contain just one term while others are longer than 32 words. An overview of the information need and the query statistics is given in Table 3.

Moreover, we wanted to contrast the queries obtained from the Yahoo! Answers with more informed queries that resemble the known-item query generation approach based on the previously suggested human computation game. For this purpose, we asked two individuals to create queries for the known movie items from the respective Wikipedia pages without using the title or actors in the query. The individuals were taught to keep the queries as close as possible to the original Yahoo! Answers question, but correcting false memories. An overview of the statistics of these queries is shown in Table 4. Compared to the figures of the queries gen-

Item	Memory in Yahoo! Answers question	Correct
Oliver & Company (movie)	[...] at the beginning there is a box of free puppies [...]	cats not dogs
Sweet And Lowdown (movie)	[...] he ends up being with a deaf woman [...]	mute not deaf
`InstantAction.com` (url)	[...] starts with an I and then the second word is games [...]	action not games

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

Yahoo! Answers information needs	
#Identified information needs	64
Average #words website known items	48.1
Minimum #words website known items	17
Maximum #words website known items	114
Average #words movie known items	77.6
Minimum #words movie known items	26
Maximum #words movie known items	194
Crowdsourcing	
#Study participants	11
Minimum #queries generated by a user	1
Maximum #queries generated by a user	64
Minimum #queries generated per known item	2
Maximum #queries generated per known item	7
Minimum #words in a query	1
Average #words per query	7.0 ($\sigma = 7.2$)
Maximum #words in a query	47

Table 3: Basic statistics of the identified information needs and the generated queries.

#Study participants	2
#queries generated per known item	2
Minimum #words in a query	2
Average #words per query	6.5
Maximum #words in a query	13

Table 4: Basic statistics of the generated queries from Wikipedia pages.

erated from the Yahoo! Answers questions, the queries from the Wikipedia pages are shorter, and their length varies less.

For the obtained queries from the Yahoo! Answers questions and for the subset of movie queries from Wikipedia pages, we evaluate the retrieval performance with respect to the ability to retrieve the desired known item. As search engines we use the API of a large commercial Web search engine as well as Carnegie Mellon's Indri ClueWeb search engine as a representative of current state-of-the-art retrieval models. The retrieval performance is measured as the mean reciprocal rank (MRR) of the desired known item in the search engines' rankings, where we only consider the first 50 results. Our rationale for considering 50 results (and not just 10) is the fact that most Yahoo! Answers users expressed a real need for re-finding the desired known item: one can assume that they would considerably look beyond the first 10 links that are typically presented to a Web searcher. Table 5 contains the results of this experiment.

Obviously, the known-item queries generated from the Yahoo! Answers questions achieve a very bad performance since none of them have the known item among the first 50 results. By contrast, the queries generated from the Wikipedia pages of the known-item movies often have the desired item among the top-5 results. This outcome clearly supports our idea that known-item queries originating from human memories are much harder (and presumably more realistic) than queries generated from the the known item itself.

5. CONCLUSIONS

As a proof of concept we have presented a new approach for the generation of realistic known-item queries. The basic idea is to use freely available questions which have a clear known-item intent, and for which examples can be found on the question answering Web site Yahoo! Answers. To ensure a wide usability in repeatable experiments, we include in our corpus only known items that are present in the ClueWeb09. In a crowdsourcing setup, human individuals are involved to generate real Web queries from the natural language questions obtained from Yahoo! Answers.

Queries generated from	Commercial Engine	Indri
Yahoo! Answers questions	0.00	0.00
Wikipedia pages	0.21	0.08

Table 5: Retrieval performance measured in mean reciprocal rank of the desired known item.

On a subset of the known-item movies we could observe that the proposed approach yields queries that are much harder for a retrieval system than the queries that are generated in a rather unrealistic setting, where humans formulate a query while having the known item before their face. Thus, we conclude that our approach resembles much better the difficulties that one encounters in the wild when working with known-item queries.

In future work, we will enlarge the corpus with the goal of using it to analyze missing and false memories that are present in known-item queries. Based on the findings, we plan to develop automatic approaches for supporting users stuck in known-item search interactions.

6. REFERENCES

[1] 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.

[2] D. Barreau and B. Nardi. Finding and reminding: file organization from the desktop. *ACM SIGCHI Bulletin*, 27(3):39–43, 1995.

[3] T. Blanc-Brude and D. Scapin. What do people recall about their documents?: implications for desktop search tools. In *IUI 2007*, pp. 102–111.

[4] R. Boardman and M. Sasse. Stuff goes into the computer and doesn't come out: a cross-tool study of personal information management. In *SIGCHI 2004*, pp. 583–590.

[5] C. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web Track. In *TREC 2009*.

[6] N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the TREC-2005 Enterprise Track. In *TREC 2005*.

[7] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. Robbins. Stuff I've seen: a system for personal information retrieval and re-use. In *SIGIR 2003*, p. 72–79.

[8] D. Elsweiler, M. Baillie, and I. Ruthven. Exploring memory in email refinding. *ACM TOIS*, 26(4):1–36, 2008.

[9] C. Hauff and G.-J. Houben. Cognitive processes in query generation. In *ICTIR 2011*, pp. 176–187.

[10] I. Hyman Jr, T. Husband, and F. Billings. False memories of childhood experiences. *Applied Cognitive Psychology*, 9(3):181–197, 1995.

[11] J. Kim and W. B. Croft. Retrieval experiments using pseudo-desktop collections. In *CIKM 2009*, pp. 1297–1306.

[12] J. Kim and W. B. Croft. Ranking using multiple document types in desktop search. In *SIGIR 2010*, pp. 50–57.

[13] H. Roediger, J. Jacoby, and K. McDermott. Misinformation effects in recall: Creating false memories through repeated retrieval. *Journal of Memory and Language*, 35:300–318, 1996.

[14] S. Tyler and J. Teevan. Large scale query log analysis of re-finding. In *WSDM 2010*, pp. 191–200.