Cognitive Processes in Query Generation

Claudia Hauff* and Geert-Jan Houben

WIS, Delft University of Technology, Delft, the Netherlands
{c.hauff,g.j.p.m.houben}@tudelft.nl

Abstract. *Known-item* search is the search for a specific document that is known to exist. This task is particularly important in Personal Information Management (PIM), where it is the most common search activity. A major obstacle to research in search technologies for PIM is the lack of publicly accessible test corpora. As a potential solution, pseudo-desktop corpora and automatic query generation have been proposed. These approaches though do not take the cognitive processes into account that take place when a user formulates a re-finding query. The human memory is not perfect, and many factors influence a user's ability to recall information. In this work, we propose a model that accounts for these cognitive processes in the automatic query generation setting.

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] and CLEF[2], have been set up over the years. They provide researchers with standardized test corpora and retrieval tasks.

While we now have access to, among others, newspaper and Web corpora, test corpora for Personal Information Management (PIM) research are still lacking due to privacy concerns. PIM is concerned with the acquisition, storage, organization and the retrieval (re-finding) of information collected by a user. Due to the ever increasing reliance on digital communication channels and functions (email, chat, etc.) as well as digitally available information, the amount of data a user stores is growing continuously. A stored item can be, for instance, an email in the user's inbox, a scientific paper the user downloaded from the Web, or a calendar entry. Re-finding an item that the user has accessed before, a process known as *known-item* retrieval, is the most common search activity in PIM. Note, that known-item retrieval is also a usage scenario of Web search engines, which users may rely on to re-find previously visited web pages [1].

Research in PIM related search technologies is hampered significantly by the lack of public test corpora. To alleviate this problem, automatic [2, 16] and human

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

[1] Text REtrieval Conference `http://trec.nist.gov/`

[2] Cross Language Evaluation Forum `http://www.clef-campaign.org/`

computation game [17] 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 topics by selecting terms of the document in question according to particular rules; for example, the most discriminative terms are selected with a higher probability or randomly (noise). In the human computation game scenario, the document in question is shown to human study participants who create queries with the goal to return the item as high in the retrieved ranking as possible. Note, that the human participants are actually shown the document, they do not need to remember it.

These two 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. Indeed, research into so-called *false memories* is an important field of study in psychology where it is often motivated by the question of eyewitness reliability [7, 22] and the correct recall of childhood experiences [12, 20]. In this paper, we argue that for known-item retrieval to be more realistic, topic generation approaches need to take into consideration the imperfection of human memory and the tendency to create false memories. A similar argument was already made by Lansdale [19] who believed that the cognitive abilities of users need to be taken into account in the design of PIM tools. 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 [11]. Based on these findings, we propose a query generation model that includes false memories in order to generate more realistic queries.

If the imperfections of human memory are not reflected in a PIM test corpus, 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 [8] 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.

The main contributions of our work are (i) an argument for the inclusion of false memories into test corpora for known-item tasks that is based on psychology research, (ii) a model for automatic query generation that includes a false memory component, and, (iii) an investigation into the TREC Enterprise track 2005 and the influences of false memories in it.

The rest of the paper is organized as follows: Sec. 2 describes research in false memories, both in psychology and PIM. Sec. 3 describes the inclusion of false memories into an existing query generation procedure. Experimental results are presented in Sec. 4, followed by the conclusions in Sec. 5.

2 Related Work

False Memories: A particular type of experiment, the Deese-Roediger-McDermott (DRM) paradigm [23], is widely used in psychology to study the effects of false memories (or memory illusions, memory distortions). A false memory is a person's recall of a past experience which differs considerably from the true course of events [24]. The DRM setup is as follows: given a critical word (e.g., *foot*) a list of no more than 15 semantically related words is created (e.g., *shoe*, *hand*, *toe*). Subjects first study the list of related terms (without the critical term), and are then asked to freely recall the terms in the list without resorting to guessing (this occurs immediately after having studied the list). Routinely, it is observed that the subjects recall the critical term, which is the elicited false memory, with a similar probability as the terms on the list. It is also notable, that study subjects are confident about having studied the critical term. One theoretical explanation for this observation has been provided by the *Source Monitoring Framework* [15, 13] (SMF), which postulates that false memories are created because of confusions about a memory's source. A source can either be internal (thinking of *foot* while having heard the terms in the list) or external (the experimenter said *foot*).

According to the SMF, a memory's source is not directly encoded in memory, instead a number of memory characteristics are exploited in order to determine the source when retrieving a memory: sensory information (sound, color), contextual information (location, time), semantic detail, affective information (the emotional state), and evidence of cognitive operations (records of organizing the information). This means for instance, when a person recalls if he has read a statement in an email, heard it from a colleague, saw it during a presentation of a talk, or thought of it himself, attributing the source will depend on the person recalling the voice of the attributor, the color of the presentation, the time of reading the email or the thought process that lead to the statement. The amount of detail remembered for each memory characteristic determines which source the person finally attributes the statement to.

Source confusion or misattribution is deemed as the main cause of false memories. Source confusion occurs when the experience is poorly encoded into memory, for instance, if somebody reads an email while being distracted by a phone call or someone walking into his office. Later, a correct recall of the email content will be more difficult than if the person would have concentrated on just reading it. Stress, distractions and a strong emotional state [14] also degrade the encoding process. When retrieving from memory, these factors influence the ability to attribute the source correctly as well. Thus, false memory attributions can be based both on the encoding and the decoding phase. Moreover, if encoded memories have largely overlapping characteristics, source confusion is more likely; recalling the differences between the memories will be difficult, while remembering the general similarities, or the gist of the memories, is easier.

While SMF explains why subjects in the DRM experiments falsely recall the critical item (they confuse thinking and reading/hearing it), the activation and monitoring theory [23] explains why they think of the critical item in the first place when hearing semantically related terms. When hearing the list terms, the memories of these terms are activated which in turn also leads to the activation of related memories (such as the critical item).

Another finding of memory research is that, the gist of a document, i.e., the meaning of the content, is longer retained in memory than specific details [25, 18]. With respect to generating topics for known-item search this means, that we need to take the amount of time passed since the user last viewed the document, or more generally the access pattern of the document to be re-found, into account. An additional factor to consider is age. It has been shown that older adults are more susceptible to false memories than younger adults [21, 9].

If we translate those findings to PIM search tools, we can argue that a PIM search system should be adapted to each individual user and the context. For instance, a PIM search system can take the age of a user into account and treat queries posed by older users differently from queries posed by a younger adult. Similarly, if the PIM search system has an indication that the user is stressed or tired (an indication may be be derived from the user's activities on the system within the last hours), a posed query may be treated differently than a query posed by a calm and relaxed user.

Personal Information Management: Blanc et al. [5] 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 when being asked to recall the title and keywords of the document in question were most often mixing true and false memories; for 32% of the documents the recalled keywords were correct, while for 68% they were only partially correct ("partial recall" in [5]). Recalling the title was more difficult: 33% correctly recalled document titles, vs. 47% partially correct and 20% completely false recollections.

Elsweiler et al. [11] 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. No indication was given if the memories were (partially) false or correct. Another finding, in line with research in psychology, was that memory recall declines over time, that is, 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 [10], where up to eight year old documents were sought by users in a work environment.

In general it has been found across a range of studies, e.g., [3, 6, 5, 4, 26], that in PIM re-finding, users prefer to browse to the target folder and to visually inspect it 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. For this reason we propose the inclusion of false memories into generated known-item queries, to make the test corpora more realistic and more in line with true user queries.

3 Methodology

In this section, we will first introduce the two types of false memories that we distinguish, based on an information retrieval point of view. Then, the automatic topic generation process, proposed in [2, 16], is briefly described before we introduce our adaptation which+ takes false memories into account.

Types of False Memories and System Responses: Recall, that in the DRM experiment (Sec. 2), the elicited false memories are semantically closely related to the true memories, as a result of the experimental setup. This type of false memories (we denote it with FM_R) can be addressed by retrieval mechanisms that add related terms to a query (e.g., synonym-based expansion, rule-based expansion, pseudo-relevance feedback). If a user searches in his emails with the query "John Saturday meeting" and the email in question contains the term "weekend" instead of "Saturday", the email can be found by such mechanisms.

While this type of false memories does not render retrieval systems ineffective, false memories that lead to a wrong recollection of the nature of the content (we denote this type with FM_F) pose a far more serious problem. For instance, the user might query the system with "John Monday meeting" or "Paul Saturday meeting"; here, the user either incorrectly remembers the time or the person he is going to meet, maybe because the user confused two meetings with each other or remembered the sender of the email, sent a long time ago, incorrectly. In these cases, current retrieval systems are likely to fail or retrieve the correct item at a low rank. Such queries do not (or very rarely) occur in the available known-item topic sets. At the same time, they are likely to occur to some extent in the real-world setting and thus they should be included in topic sets that are utilized to test and evaluate PIM retrieval systems.

Automatic Topic Generation: The known-item topic generation approach originally proposed by Azzopardi et al. [2] was later refined by Kim et al. [16] for the more specific case of PIM test corpora, where a document usually contains a number of fields (such as email sender, calendar entry time, Word document creator, etc.). A known-item/query pair is then generated in five steps:

1. Initialize an empty query $q = ()$
2. Select document d_i to be the known-item with probability $P_{doc}(d_i)$
3. Select the query length s with probability $P_{length}(s)$
4. Repeat s times:
   (a) Select the field $f_j \in d_i$ with probability $P_{field}(f_j)$
   (b) Select the term t_k from field language model of f_j: $P_{term}(t_k|f_j)$
   (c) Add t_k to q
5. Record d_i and q as known-item/query pair

Kim et al. [16] verified that this query generation procedure is more similar to queries generated in their human computation game than queries generated without considering the separate fields. In their work, P_{term} is based only on the target document, that is, no noise is included in the query generation process. In contrast, Azzopardi et al. [2] proposed to interpolate P_{term} with random noise from the background model (collection language model) to simulate a user with an incomplete recollection of the content. If applied to fields, the term selection probability becomes:

$$P_{term} = \alpha P_{term}(t_k|f_j) + (1-\alpha)P(t_k), \tag{1}$$

where $P(t_k)$ is the probability of drawing t_k from the background model of the respective field. The probabilities P_{field}, P_{doc}, P_{term} and P_{length} can be chosen in a number of ways. Following the experiments in previous work, we draw fields uniformly at random [16], we draw the query length s from a Poisson distribution [2], and rely on TF.IDF based term selection. The TF.IDF based term selection has been shown in [16] to lead to generated queries that are more similar to manually created (TREC) queries than other approaches.

Modelling False Memory: Based on Eq. 1, a first step is to make the parameter α dependent on the time the known item was last seen, instead of fixing it to a particular value across all documents. This step can be motivated by the increase of false memories over time: if a document has not been seen in a year, a user is more likely to have a false memory of it compared to a document last viewed the day before.

Let x_{d_i} be the number of time units since document d_i was last seen and let x_{max} be a time unit where no document specifics are remembered anymore (and $x_{d_i} \leq x_{max}$), then we can model α as follows:

$$\alpha_{d_i} = \left(\frac{x_{max} - x_{d_i}}{x_{max}}\right)^n, \alpha \in [0,1] \; and \; n > 0 \tag{2}$$

If document d_i has recently been viewed α_{d_i} will be ≈ 1 and little noise is introduced in the query generation process. On the other hand, if a document has not been viewed for a long time, α_{d_i} will be ≈ 0 and a large amount noise is introduced. The parameter n determines how gradual or swift the introduction of noise is over time: the closer n is to 0, the more gradual the memory loss; conversely, the greater n, the quicker the introduction of noise. Adapting the level of noise to the access pattern of the target document is not the only possibility. In Sec. 2 we described how numerous factors (stress, emotional state, context, etc.) affect the encoding and decoding of a memory and if those factors can be measured, they should influence the noise level as well.

We have stated earlier, that random noise (terms drawn from the collection) is not a realistic modelling decision, as users are likely to retain some sense of what the document they look for is about (e.g., a meeting with some person on some day). Recall how in Sec. 2 we discussed the source monitoring framework which has been proposed and empirically validated as an explanation of false memories. Based on it, we model the noise (false memories) as coming from different sources $S_1, .., S_m$. One source may be constructed from the documents semantically related to the known item, another source may be derived from all emails sent by a particular sender, and so on. External sources may also be utilized as source, e.g., news stories that were published at the time the target document was received/read/sent.

As a consequence, we adapt step 4(b) in the query generation process to include levels of noise that are dependent on the amount of time passed since the document was last seen by the user and to draw noise from a number of sources that are related to the target document:

$$P_{term} = \alpha_{d_i} P_{term}(t_k|f_j) + (1-\alpha_{d_i}) \left(\sum_{\ell=1}^{\ell=m} \beta_\ell P_{term}(t_k|S_\ell) \right),\ with \sum_{\ell=1}^{\ell=m} \beta_\ell = 1 \quad (3)$$

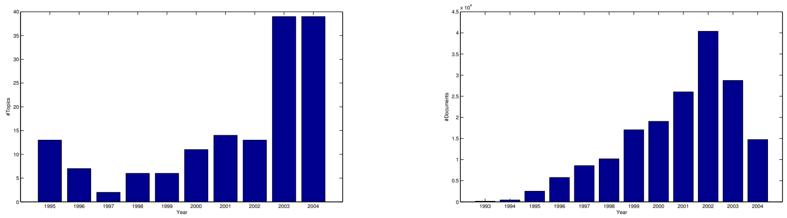
4 Experiments

As PIM test corpora are not publicly available, we consider instead the email corpus (W3C corpus) introduced at the TREC Enterprise track in 2005 [8]. The Enterprise track was developed with the question in mind of how people use enterprise documents (intranet pages, emails, etc.) in their workplace. One of the tasks was the re-finding of emails, which is the task we investigate here. We consider it a reasonable approximation of a PIM search corpus and note that it was also utilized in previous Desktop search experiments [16].

Data Set Analysis: The W3C corpus contains (among others) $198,394$ email messages from the public mailing list *lists.w3.org*. A total of 150 topics were developed (25 for training and 125 for testing) by the task participants. Though no detailed information is given in [8] concerning the topic creation process (the topics were created by the participants), it can be assumed that the task participants viewed the email messages while developing the topics.

A total of 67 runs were submitted to TREC in 2005 for the email re-finding task. The retrieval effectiveness was measured in mean reciprocal rank (MRR) and success at 10 documents (S@10). The best system achieved a performance of 0.62 (MRR) and 82% (S@10). The task was not further developed in the following years; the performance of the best systems appeared to indicate that known-item search in such an email corpus is not a difficult problem. In the subsequent paragraphs we show that this conclusion can only be drawn if we assume the existence of perfect memory.

In Sec. 2 we described studies that have shown that memory degrades over time. An obvious question is then, how distributed are the documents in this corpus and the 150 target documents (qrels) over time. In Fig. 1 we present histograms (in years) across all corpus documents and the relevant documents only. The documents cover a ten year time span, from 1995 to 2004. While most relevant documents are from 2003 and 2004, more than ten known items are emails written in 1995. If we assume (due to a lack of user logs to investigate actual document access patterns) that the documents were read once they were received, it becomes clear that perfect queries for those documents is an unreasonable assumption.



(a) Qrels documents

(b) Corpus documents

Fig. 1. Histograms of the number of documents according to year of sending.

The query generation process in Sec. 3 takes the fields of a document into account. From the corpus we extracted the following fields: *sender*, *subject*, *body* and *sending date*. We then manually assessed the 150 topics and assigned their terms and phrases to one or more of the fields. This assessment evaluated false memories of type FM_F : if the query terms match the subject line (or email body, sender, date) semantically, the terms are judged as being correct memories, even if not all terms occur as such in the emails. If a query's terms are applicable to several fields, e.g., subject and body, they are assigned to all applicable fields. Query terms/phrases are deemed a FM_F false memory if they are false in the context of the target email document. For instance, topic *KI6* (Fig. 2) is: *Conference on accessibility and assistive technology at schools*; the known-item specifically discusses a conference on assistive technologies for colleges and universities, not schools; this topic thus contains a FM_F false memory. Due to the topic construction process, we expect very few topics to contain false memories, which we argue is in contrast to real-world queries.

```
<annotatedTopic>
<num>KI6</num>
<qrel>lists-076-5352080</qrel>
<originalEntry>Conference on accessibility
and assistive technology at schools</originalEntry>
<sender></sender>
<date></date>
<subject>Conference assistive technology</subject>
<body>Conference on accessibility and assistive technology at</body>
<falseMemory>schools</falseMemory>
</annotatedTopic>
```

Fig. 2. Topic annotation example (FM_F).

	FM_F	FM_R
Field	**#Topics**	**#Topics**
sender	23	22
date	12	17
subject	32	129
body	147	132
false memory	14	51

Fig. 3. Number of topics containing information present in a field.

We also performed this topic set analysis automatically, focusing on false memories of type FM_R , that is, we considered the syntactic matching between query terms and document terms. The email corpus and the topics were stemmed (Krovetz) and stopwords were removed[3]. Here, a topic contains a false memory, if at least one of the query terms does not occur in the email document.

In Tab. 3 the results of this analysis are shown. For both types, FM_F and FM_R , the vast majority of topics contain elements from the subject and/or the email body. Few topics contain additional aspects such as the sender or the date of

[3] All retrieval experiments were performed with the Lemur Toolkit: `http://www.lemurproject.org/`

sending. While the number of false memories is low in the FM_F setting, about a third of emails contain false memories of type FM_R .

In order to investigate how those false memories influence the performance of retrieval systems, we evaluated all 67 runs[4] submitted to TREC in 2005 on the four subsets of topics: (i) the topics without FM_F false memories, (ii) the topics with FM_F false memories, (iii) the topics without FM_R false memories, and, (iv) the topics with FM_R false memories. The question is: Do the same runs that perform well on topics without FM_R or FM_F false memory topics also perform well on the topics with these false memories? The results are shown in Fig. 4. Plotted are the system performances in MRR: the performance on topics without FM_F / FM_R false memories (x-axis) versus the performance on topics with false memories (y-axis). Fig. 4 (left) shows the scatter plot for the topic split according to FM_F and Fig. 4 (right) shows the topic split according to FM_R . We are interested in how similar the system rankings are. Ideally, the system rankings would be the same independent of the topic set. This is not the case, in fact, the rank correlation between the two sets of system performances for the FM_F based topic split is not statistically significantly different from zero (at $p < 0.01$). In contrast, for the FM_R based topic partition, the correlation is significant and a trend is recognizable (Fig. 4 (right)). However, even here the best retrieval systems across *all* topics do not fare well. The best system across all topics is placed at rank 27 of the FM_F topics, while it is ranked ninth in the FM_R topics. In case of the correct topics, the best system is within the top five ranks, both for the topic partition without FM_F and without FM_R false memory topics. This result shows, that systems that perform well on one type of topics (topics without false memories) may perform rather poorly on topics with false memories; a factor that needs to be taken into account when researching retrieval approaches in PIM. This result also emphasizes the need for more realistic queries, i.e., those with realistic false memories.

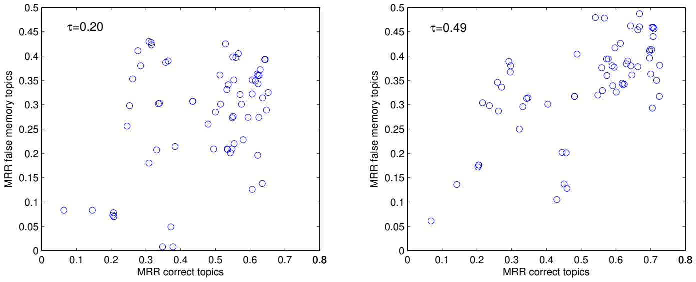


Fig. 4. Scatter plots of system performances (in MRR): on the left, the topics without FM_F false memories (x-axis) are plotted against the topics with FM_F false memories (y-axis). On the right, the topics without FM_R false memories are plotted against the topics with FM_R false memories.

[4] The runs are available at `http://trec.nist.gov/`

Query Generation with False Memories: In this section, we report the results of our query generation approach and its influence on three standard retrieval approaches: TF.IDF, Okapi and Language Modeling with Dirichlet smoothing ($\mu = 1000$). As source S of false memory for a field f_j of the known-item document d_i, we utilize the 1000 most similar fields (cosine similarity) of f_j in the corpus. We evaluate two decay rates, $n = \{1, 2\}$. Finally, we derive topic sets, each of size 100, which contain known-items of different sending date (the "current date" is the day of the most recently correctly time-stamped document in the W3C corpus). The derived topic sets are:

- **Random**: the known-item documents are drawn at random from the corpus; their distribution of document age (document sending date) will resemble Fig. 1.
- **Cold**: the known-item documents were not sent within the last year.
- **Warm**: the known-item documents were sent between a year and three months ago.
- **Hot**: the known-item documents were sent within the last three months.

Topics that belong to the "hot" (recently seen) category contain the smallest amount of noise, while topics in the "cold" (not seen for a long time) category are highly likely to contain a lot of noise (Eq. 2). The noise-controlling parameter α_{d_i} is determined for each known-item document d_i by calculating the fraction of years that have passed since the document's creation (x_{d_i}); x_{max} is set to 10 years (the time interval of the corpus). The results are presented in Tab. 1. The worst results are recorded for "cold" queries, which is not surprising as they were generated with the most noise. In general, the results confirm the expectations, no single retrieval approach performs best overall. The absolute performance changes drastically between the hot and cold query sets, indicating the suitability of the model to introduce false memories.

Ideally, we would like to compare the generated queries to an existing topic set (as done in [16]), to investigate the model's ability to generate queries and false memories that are similar to manually created queries and naturally occurring false memories. This is, however, not yet possible, as no known-item topic set exists, which includes topics that were created in a realistic setting.

5 Conclusions

In this work, we have argued for taking cognitive processes into account when generating queries, in particular queries in the PIM setting and the known-item task. We have shown experimentally, that false memories can have a significant impact on the relative performance of retrieval systems and we proposed a false memory based adaptation of the existing query generation procedure.

A limitation of our work is the adhoc nature of the parametrization, e.g., we sampled known items uniformly from the corpus or according to a certain time-stamp range, though it would be very useful to know when the documents, that users typically search for, were last seen by them. In order to compare how well

Query Set	n	**TF.IDF**	**Okapi**	**Dirichlet LM**
Random	1.0	0.465	0.443	0.493
	2.0	0.311	0.368	0.390
Cold	1.0	0.249	0.260	0.251
	2.0	0.234	0.255	0.255
Warm	1.0	0.671	0.690	0.597
	2.0	0.583	0.587	0.596
Hot	1.0	0.701	0.713	0.777
	2.0	0.566	0.699	0.679

Table 1. Results of known-item retrieval (in MRR) for generated query sets with different sending date characteristics.

our model approximates the true amount and type of false memories in re-finding queries, we need to collect re-finding queries from real users. To that end, we plan to follow the following two approaches:
(1) In [16] it is argued that the introduced pseudo-desktop corpus is valuable, because the users who played the human computation game were already familiar with the documents (e.g., e-mails sent through a university mailing list). Instead of letting users "play a game" to find the best possible query, we plan to ask a set of users about such publicly accessible e-mails without letting them view the document. Choosing documents that were sent across a wide time span, will give an indication of how large the false memory problem is in this setting. A potential pitfall is here to direct the user to the right document, without biasing the keyword search through the description.
(2) False memories can also be observed in newsgroups and discussion fora. A typical post in a newsgroup or a forum may be: *"I know that I saw a post about this another time, explaining where to find the program in order to uninstall it, but I cannot find the post. Can someone send me a link to that post, or post the information again please?"* and one or more of the replies then point to the original post the user was looking for (confirmed by an affirmative statement of the original requester). These are also false memories in a known-item setting: a user is certain that an item exists, but he cannot find it. The posting dates of the different entries also allows an investigation into false memories over time.

References

1. E. Adar, J. Teevan, and S. Dumais. Large scale analysis of web revisitation patterns. In *SIGCHI '08*, pages 1197–1206, 2008.
2. L. Azzopardi, M. de Rijke, and K. Balog. Building simulated queries for known-item topics: an analysis using six european languages. In *SIGIR '07*, pages 455–462, 2007.
3. D. Barreau and B. Nardi. Finding and reminding: file organization from the desktop. *ACM SigChi Bulletin*, 27(3):39–43, 1995.
4. O. Bergman, R. Beyth-Marom, R. Nachmias, N. Gradovitch, and S. Whittaker. Improved search engines and navigation preference in personal information management. *ACM Trans. Inf. Syst.*, 26(4):1–24, 2008.

5. T. Blanc-Brude and D. Scapin. What do people recall about their documents?: implications for desktop search tools. In *IUI '07*, pages 102–111, 2007.
6. 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 '04*, pages 583–590, 2004.
7. Q. Chrobak and M. Zaragoza. Inventing stories: Forcing witnesses to fabricate entire fictitious events leads to freely reported false memories. *Psychonomic bulletin & review*, 15(6):1190–1195, 2008.
8. N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the TREC-2005 Enterprise Track. In *Proceedings of TREC 2005*, 2005.
9. C. Dodson, S. Bawa, and S. Slotnick. Aging, source memory, and misrecollections. *Learning, Memory*, 33(1):169–181, 2007.
10. 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 '03*, pages 72–79, 2003.
11. D. Elsweiler, M. Baillie, and I. Ruthven. Exploring memory in email refinding. *ACM Trans. Inf. Syst.*, 26(4):1–36, 2008.
12. I. Hyman Jr, T. Husband, and F. Billings. False memories of childhood experiences. *Applied Cognitive Psychology*, 9(3):181–197, 1995.
13. M. Johnson, S. Hashtroudi, and D. Lindsay. Source monitoring. *Psychological Bulletin*, 114(1):3–28, 1993.
14. M. Johnson, S. Nolde, and D. De Leonardis. Emotional focus and source monitoring. *Journal of Memory and Language*, 35:135–156, 1996.
15. M. Johnson and C. Raye. Reality monitoring. *Psychological Review*, 88(1):67–85, 1981.
16. J. Kim and W. B. Croft. Retrieval experiments using pseudo-desktop collections. In *CIKM '09*, pages 1297–1306, 2009.
17. J. Kim and W. B. Croft. Ranking using multiple document types in desktop search. In *SIGIR '10*, pages 50–57, 2010.
18. W. Kintsch, D. Welsch, F. Schmalhofer, and S. Zimny. Sentence memory: A theoretical analysis. *Journal of Memory and Language*, 29(2):133–159, 1990.
19. M. Lansdale. The psychology of personal information management. *Applied Ergonomics*, 19(1):55–66, 1988.
20. E. Loftus and J. Pickrell. The formation of false memories. *Psychiatric Annals*, 25(12):720–725, 1995.
21. K. Norman and D. Schacter. False recognition in younger and older adults: Exploring the characteristics of illusory memories. *Memory & Cognition*, 25(6):838–848, 1997.
22. 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.
23. H. Roediger and K. McDermott. Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology-learning memory and cognition*, 21(4):803–814, 1995.
24. H. Roediger III and K. McDermott. False perceptions of false memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:814–816, 1996.
25. J. Sachs. Recopition memory for syntactic and semantic aspects of connected discourse. *Attention, Perception, & Psychophysics*, 2(9):437–442, 1967.
26. J. Teevan, C. Alvarado, M. Ackerman, and D. Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *SIGCHI '04*, pages 415–422, 2004.