Exploring Query Patterns in Email Search

Morgan Harvey and David Elsweiler

[1] Dept. Computer Science 8 (AI), Univeristy of Erlangen-Nuremberg, Germany.
[2] Institute for Information and Media, Language and Culture, University of Regensburg, Germany
morgan.harvey@i8.informatik.uni-erlangen.de, david@elsweiler.co.uk

Abstract. Despite Email being the most popular communication medium currently in use and that people have been shown to regularly re-use messages, very little is known about how people actually search within email clients. In this paper we present a detailed analysis of email search behaviour obtained from a study of 47 users. We uncover a number of behavioral patterns that contrast with those previously observed in web search. From our findings, we describe ways in which email search could be improved and conclude with a short discussion of possible future work.

1 Introduction

Despite the recent explosive growth of social media applications, email remains the most popular communication medium in use today. 92% of online American adults use email [11] and an estimated 294 billion emails are sent each day [12][3]. Email is, however, much more than just a communication tool. People use email for diverse purposes including the management of tasks, projects, contacts and content[17]. Email is also not the ephemeral media it was originally intended to be. Most messages have lifespans of several weeks or months and some messages are re-read years after they were first received [7].

Reflecting the diverse usage patterns and long-life spans, studies of desktop search logs show that email messages tend to be searched for more often than any other kind of media, including visited web pages [4]. There is also evidence that searching emails can be often be difficult and time consuming [5]. Nevertheless, in contrast to other media, and web pages in particular, search behaviour for email messages has received little research focus.

In this paper, we address this situation by analysing the queries and resulting clicks of 47 Mozilla Thunderbird users over the course of a 4 month period. We examine several features of querying behaviour including how people resubmit the same or similar queries over time and how click-through patterns change in different situations. In doing this analysis we uncover a number of useful patterns and behaviours in query usage that can inform the future development of email clients.

[3] This compares to 1 billion facebook entries [1] and 200 million tweets [2] per month

2 Related Work

For over a decade search engine (SE) log analyses have been the primary method for learning about how people search. Early analyses focused on snapshots of behaviour in short time windows in such logs. These analyses have provided valuable information characterising user queries and sessions, e.g. [14, 10], and give a useful overview of search behaviour on the Web.

More recent work has looked at querying behaviour over significantly longer time periods; examining temporal aspects such as query repetition and how pages are re-found [13, 15, 16, 3]. These analyses reveal that query re-use behaviour is extremely common. For example 33% of all SE queries submitted have been identically submitted previously by the same user and for 39% of all queries the user returns to a Web page that he has found before via a separate search [15]. SE log analysis has shown that queries submitted with the intention of re-finding are typically shorter than those for new content and the clicked on pages for such queries tend to rank higher in the results list [16].

Sanderson and Dumais [13] looked at how SE queries are repeated over time and found that individual users are very likely to repeat the same query for around 7 days, after which point the probability of re-use tails off rapidly. Further, they observed different fall off rates for different types of query. Navigational queries (where the goal is to find a particular web site), for example, tend to be repeated over longer periods of time than non-navigational queries. Adar and colleagues [3] also examined temporal patterns in the logs, but focused on repeat visits to web pages over time. They observed 4 clear patterns of re-visitation and discovered that the pattern will depend on several factors including a person's intent, page content and site structure. These studies have provided important insights into how search engines should be designed to support certain behaviours in different situations, in order to provide a better experience for users.

The large and varied body of literature available on web search contrasts with that of email search behaviour, which has mainly been studied in the context of desktop search. Desktop search queries are typically much shorter than web queries and often contain named entities [4]. However, desktop search queries only account for a fraction of email searches. The only lab-based study of email re-finding in the literature found similar trends [7] with approximately 60% of queries containing a reference to named entities and 40% containing a reference to a person. In our previous work we reported on a naturalistic study of email re-finding behaviour [6]. This work revealed several important aspects of email search behavior. It confirmed previous findings regarding the frequency of searching and showed that users often experience difficulties. Search efforts regularly contain large numbers of queries, message clicks and can last for long time periods. Further, many of the search attempts involved the user clicking on the same messages or same folders multiple times, indicating user disorientation.

The work presented in this paper builds on these first analyses to look at query patterns over time by applying similar techniques to those that yielded such useful results for web search. In doing so we find out how email behaviour differs from that on the web and reason about why and how these differences

should affect the analysis of such data and how it might inform email-client design.

3 Data Collection

In order to study email search it is necessary to obtain a sufficiently large log of interactions of a variety of users with an email client. Unlike web search - where logs can be collected by search engine companies with millions of users - this requires a user study where software written to collect such log data is installed within an email client over a period of time. Our data were collected by conducting a naturalistic study of email use with the popular, open-source email-client Mozilla Thunderbird[4].

We developed and deployed a custom software extension that recorded user interactions with the client including messages that were read, clicks on folders, clicking on column headers to sort mails and search queries submitted. Full details of the data collected can be found in our previous publication [6]. 47 participants with diverse backgrounds (37 male, 10 female, aged 21-49, from 7 countries) volunteered to take part over a period of 4 months.

3.1 Deriving Query Chains

The basis of our analyses in this paper are what we refer to as query chains, which were created by associating each query with messages that were subsequently clicked. Chains are ended when there is a gap between message clicks of 5 minutes or more (as has been used before in desktop search), a new query or a re-start of the email-client. We chose to analyse query chains as our previous work found the overwhelming majority of re-finding attempts with an email-client started with a search query [6] and therefore they serve as good proxies for re-finding behaviour. A second benefit is that this gives us a dataset in a similar form to search engine logs where each individual query is associated with a set of clicked items. Reflecting the kinds of analyses we wanted to perform, we decided to omit queries without any subsequent clicks from our dataset (41 %). We also ignore interactions, such as sorts and folder clicks. All of these aspects will be dealt with separately in future publications. Here we focus on submitted queries and subsequent clicked messages.

One issue with our data is that many of the tasks will be interwoven with other unrelated email tasks. For example, upon finding what they want a user might return the information space to its standard state and in the process messages not related to a search will be clicked or selected. Further, shortly after completing a search the user may receive and read new messages. These kinds of behaviour result in query chains with very recently received messages being logged as "clicked" at the end of chains. We chose to deal with this by removing

[4] See http://www.mozilla.org. We used version 2.3 and data was collected between June and October, 2009

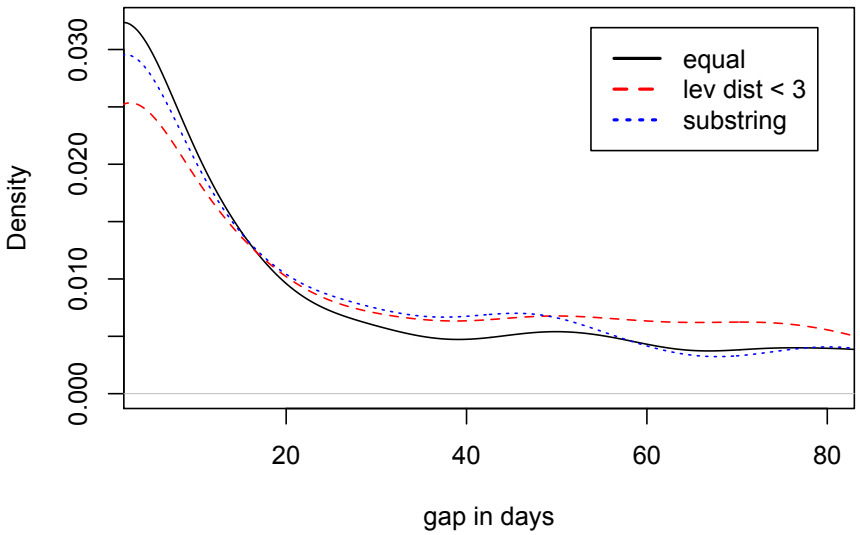

Fig. 1. Density of query re-use over time gap in days

messages that have been received in the last 48hrs from the chains. Statistical analyses of these newly received mails shows that they were significantly more likely to be at the end of chains than older messages (mean chain position 71% vs 55%), evidencing our assumption that these messages in most cases were not representative of search behaviour, but are merely an artefact of the logging process. We acknowledge that in a small number of cases this may have resulted in the removal of messages genuinely belonging to query chains, as well as genuine searches for new messages.

4 Overview of Query Chains

The process of creating query chains, as described above, resulted in a dataset containing 1467 query chains. Most (94%) of the queries were single words and only 0.56% had a length greater than 2 terms, with the mean number of terms in the queries being 1.07. The mean character length of queries was 5.87 and many were only partial words. Although in line with our previous analyses [6], this is much shorter than reported elsewhere in the literature with 1.6 words being reported for desktop search [4] and 12.1 characters for web page re-finding [16]. It is also shorter than the 1.4 words reported for lab-based studies of email re-finding [7].

4.1 Repeat Queries and Message Clicks

We were interested in examining how often people try to search for the same messages. A starting point with respect to this goal was to examine repeat queries

from individual users. In our data, 45.3% of queries were subsequently repeated by the same user, which is somewhere between the 33% [15] and 50% [13] reported for web data. 7.4% of all queries were repeated across users. Again this is very similar to the figure reported for web data. This is somewhat surprising given the very personal nature of emails in comparison to the web. Examples of queries that overlapped were common fore/surnames, names of events and software and terms that one can imagine being frequently searched, such as "deadline".

We also investigated how individual users repeated queries over time. The solid line curve in Figure 1 graphically demonstrates how this behaviour occurred, depicting the density of query re-use over time gaps in days. The density is based on a Gaussian kernel smoothed density of the histogram of time gaps binned into days with a bandwidth equal to the standard deviation of the kernel function, as is standard practice for kernel density smoothing. From Figure 1, we can see the same pattern observed by Sanderson and Dumais [13] for query resubmission for individual users with search engine logs. That is that as time goes on there is less chance of the same user re-submitting the same query. There tends to be a short period after a query is submitted in which there is a high probability that the same query will be repeated. However, the probability tails off sharply at first and then smooths off with a long tail. After around 20 days the probability remains more or less constant. We describe Figure 1 in greater detail later in the paper.

In performing the same analysis with message re-clicks we find that while the distribution does have a similar shape to those in Figure 1 it has a much quicker and sharper drop-off. In contrast with the re-use patterns for queries it seems that very few emails are re-clicked over long time gaps with only 356 of 3910 in total (9.1%) being re-clicked more than 30 days apart. As we reported previously in [6], there is evidence in our data for a small number of messages having long lifespans.

5 Repeat Behaviour Scenarios

Building on these initial analyses of query re-submission and message re-clicks over time, we wanted to use our data to gain an improved understanding of the user's intention in different re-finding situations. To this end, using our data, we created Figure 2, which is equivalent to the one Teevan et al. derived for search engine log data [15]. This approach allows us to investigate re-finding behaviour (repeated clicks and repeated queries) from 2 perspectives. We can look at click-through patterns when queries are the same or different and we can examine the queries when the clicked on messages are same, similar or different. Recreating Teevan et al's table also gives us a platform from which we can draw comparisons with previously reported search engine behaviour.

To derive the table we first define a number of concepts. We have a set of query chains Q where each chain contains a single query q_i (where subscript i is the index) and a set of email messages (click chain) $m_{i,j} \in M_i$. A query q_1 is

All query chains 1467 (100%)	**Overlapping Click Chains** - 784		**No Common** – 932 (63.5%)
	Equal Click Chains 138 (9.4%)	**Some Common** 646 (44.0%)	
Equal Query Queries 665 (45.3%)	(1) 58 (4.0%)	(2) 333 (22.7%)	(3) 274 (18.7%)
Different Query 802 (54.7%)	(4) 82 (5.6%)	(5) 313 (21.3%)	(6) 407 (27.7%)

Fig. 2. Table of refinding scenarios, replicating Teevan et al.'s [15] table

said to be equal $\iff \exists\, q_2 :\ q_1 = q_2$ and is different $\iff \nexists\, q_2 :\ q_1 = q_2$. A click chain M_1 is said to be equal $\exists\, M_2 :\ \forall m_{1,j} \in M_1 \exists m_{2,j} \in M_2 : m_{1,j} = m_{2,j}$ and overlapping $\iff \exists\, M_2 :\ M_1 \cup M_2 \neq \emptyset$. There are several specific situations of interest in the table. Below we investigate these in detail.

5.1 Click Overlap with Equal / Different Queries

Intuitively we would expect that when an equal query is submitted the subsequent message clicks would be more likely to contain message overlap than two different queries (cells 1, 2 and 3 in Figure 2). This was the situation reported in [15] for SE log data. However, our data show only slightly higher chance of overlap when the queries are the same compared with different queries. 59% (58+333 / 665) of equal query pairs had overlap while 49% (82+313 / 802) of different query pairs had some overlapping messages. The small difference between these figures could be an artefact of the way people search for emails and the properties of email collections themselves. Firstly, the query chains are much longer than for SE log data, i.e. the number of messages clicked per query is much larger. Email query chains had on average 5.17 message clicks compared to the 1 or less page clicks associated with SE queries [16]. Secondly, email collections are of course much smaller than the web, which naturally increases the chance of serendipitous overlap.

A clearer picture of what is happening in these situations can be attained by investigating the amount of overlap using overlap coefficient as a metric. Given two sets of comparable items S_1 and S_2 the overlap coefficient is calculated as follows:

$$Overlap(S_1, S_2) = \frac{|S_1 \cap S_2|}{min(|S_1|, |S_2|)}$$

The metric describes how much of the smaller set is included in the larger and is not sensitive to the relative sizes of the two sets. Applying this metric we find that there is a 78% overlap when queries were the same compared with 40% when queries were different but there was some overlap. This is a clear difference. 40% still seems high, however, but this can be explained by further investigation of the queries involved. Many of the queries that were counted as different were

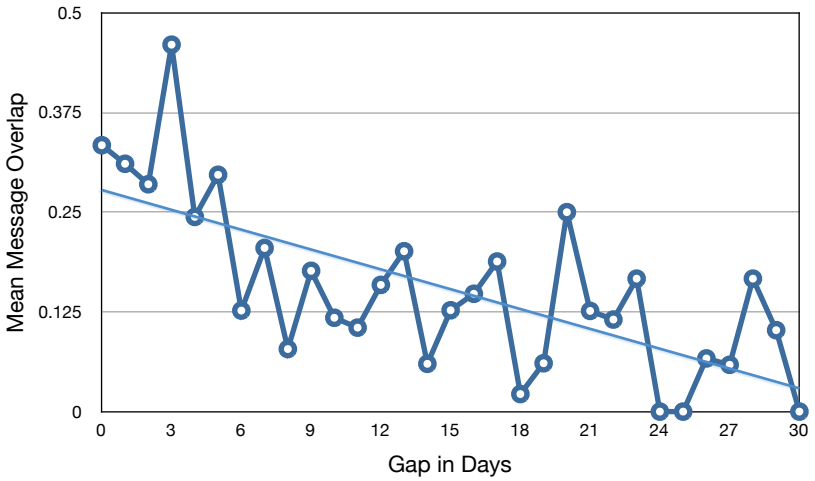

Fig. 3. Mean message overlap as time gap between equal queries increases

in fact very similar (we explore query similarity in detail below). Further, some chains had few (or even 1) message clicks and when there was overlap in these messages this distorts the mean.

5.2 Equal Query - No Overlap

We wanted to try and fully understand what is going on when a repeat query is submitted but the user clicks on a completely different set of messages, i.e. there was no overlap between the clicked on messages (cell 3 in Table 2). This happens very infrequently (4%) in search engine logs [15], but comparatively frequently in our data (18.7%). Further analysis of the clicked messages shows that although, in this situation, the users clicked on different messages, there tends to be high sender overlap between the messages clicked (39.4% of messages have the same sender) [5]. This suggests that firstly, the users are searching by sender (which aligns with what we know about the queries submitted), but also because the messages have no overlap whatsoever, it appears as though in these situations they are looking for different messages.

We wanted to investigate if there was a temporal dimension to the user intent i.e. does what the user wants when they repeat a query change with time? To achieve this we plotted percentage of click overlap for equal queries over time [Figure 3]. This figure highlights clear differences between web and email search behaviour. Teevan et al. [15] reported a trend whereby very recent re-submitted queries (with a delay of up to a few hours) had a low probability of repeated clicks. We found no evidence for this in our data. This can be explained by the fact that in web search, when people re-submit a query shortly after previously

[5] To establish sender overlap we used the same overlap coefficient metric as above

submitting it, they can be looking for different sources of information for a particular or related information need. This would explain clicks on different pages. In email searches, on the other hand, the user usually has a specific email in mind that he believes will solve his need [8]. Thus, it makes sense that this behaviour is not visible in our data.

Figure 3 also shows a clear temporal aspect to information needs. The amount of overlap in clicked on messages decreases linearly as time between query pairs grows (R^2=0.4855). If you take high message overlap to be an indicator of re-finding intent, this suggests that the information need associated with a query changes with time. This pattern contrasts sharply with the findings reported for web search where a repeated query has at least 90% probability of a repeat click when it is repeated up to 30 days after the previous submission [15]. In our data, as shown in Figure 3, when queries were repeated after a delay of 30 days there was very little overlap in the message clicks.

The reason for this is likely due to the fact that many of the email queries were people's names and if that person regularly sends emails then it is very possible a repeat query will be submitted with the intention of finding a different message. We explore popular senders in the message clicks in Section 6 below.

5.3 Overlapping Clicks With Different Queries

A final situation of interest in Table 2 is when there was overlap in clicks, but the queries were not the same (cells 4 and 5). Examining these queries however, reveals that although the queries were not the same, often they were very similar. Examples of query pairs for chains where message overlap is high (overlap coefficient >75% [6]), but queries do not match include:"mar"/"maria", "jen"/"jennifer", "virt"/"virtual", "lisa"/"lisa nathan", "johnston"/"david joh"[7]. These examples highlight that many of the similar query pairs had a small lexical change or that one query was a sub-string of the other, which seems to endorse our use of message overlap as a determiner of re-finding intent. Investigating this further, we discovered that 29.4% of time when there was high overlap (i.e. >=75%) the query was exactly the same. In 41% of high-overlap chains the Levenshtein distance between queries was 0 or 1, i.e., at most only one character different. As Table 1 shows, when queries were similar (i.e. they have a Levenshtein distance <=1) then the messages clicked were also very similar. Queries that had a distance of >=2 had, however, very little click overlap. This is further evidence in favour of our assertion that high query overlap is an indicator of re-finding intent.

Figure 1 shows query re-use over time. This is shown for equal queries (solid) and queries deemed similar due to them either having small Levenshtein distance (dashed) or one being a sub-string of the other (dotted). All three density curves suggest that there is a frequent need to re-find the same message within about

[6] 103 chains have message overlap >75% with at least one other chain, but do not have complete overlap

[7] Names have been changed for privacy reasons

Levenshtein Distance	Message Overlap Coefficient (%)
0	17.65
1	14.56
2	2.65
3	0.51
4	0.29
5	0.21

Table 1. Message overlap against Levenshtein distance between queries

a week, but this tails off after time. As the curves for similar queries tail off at a slower rate, this suggests that people are trying to re-find the same things longer than the equal queries line suggests, but they are less likely to recreate exactly the same query.

This means that we can use two clues to understand the user's intent when they submit a repeated or similar query. First, how long has it been since they last submitted that query. The longer the time period elapsed since the query was last submitted, the less chance there is that they are looking for the same message. Second, the similarity of the query submitted to a previously submitted query. The closer the lexical similarity, the higher the chance of clicked message overlap, which we take as a strong indicator of re-finding intent.

6 How do people feature in the search results?

We know from the literature (and our query data) that many email queries are references to a person. We also know from the analyses above that there is high sender overlap in the click-through patterns. We wanted to examine if particular senders were important, in particular we wanted to look at the people who most frequently send our participants emails and investigate how often these individuals feature in the click-through patterns. To do this we examined how frequently the top k senders (i.e. top 1 sender would be the sender who most frequently sends an individual participant mails) for each participant in the study appear in the messages clicked.

Overall, top k senders feature extremely regularly in the click-through data. 40% of all query chains contain at least 1 top 5 sender. Further, 25% of all messages in query chains are from top 5 senders with 16% of messages in the chain having been sent by the top sender. Figure 4(left and right) show the relationship between top k senders in the collection and their presence in the click through data graphically. The chart on the left shows the percentage of query chains that contained at least one message sent by one of the top k senders, along with the estimated probability of such an event. The chart on the right shows the percentage of messages in chains that were sent by the top k senders and the probabilistic expectation. Both of the probability estimates were calculated based on the ratio of emails received that were sent by the top k senders over the total messages received. For the message percentage case this is a simple estimate, however for the other case this is obtained by calculating the probability of observing that sender at least once given n draws from the multinomial

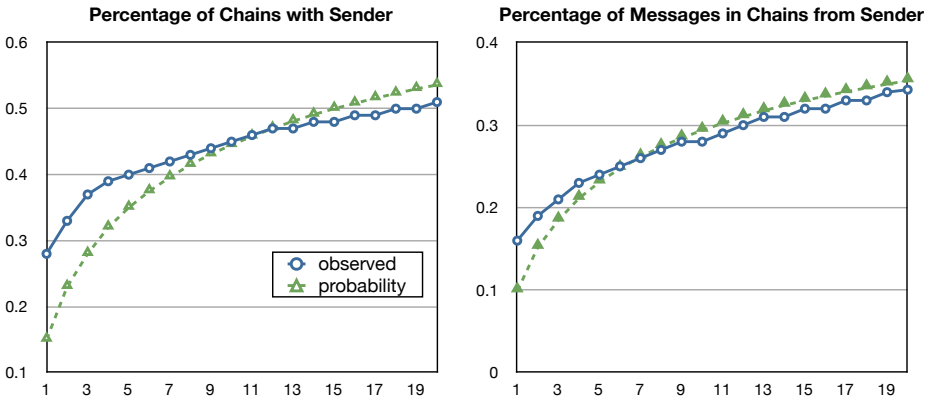

Fig. 4. Top k senders

distribution over senders, where n is the number of messages in the chain. Note that this is the same as the complement of not observing that sender over the n trials. These figures show that the top 10 senders and the top 5 in particular were searched for far more often than would be expected based on the frequency with which they send mails alone. Messages from these frequent senders also make up a far larger percentage of clicked messages than would be expected by chance. Interestingly, on both charts there is a point between k=6 and k=10 where messages from senders tended to be clicked on less than would be expected based on the number of messages from them in the collection.

7 Summary of main findings

To summarise, we have analysed the query and corresponding message click data from a naturalistic study of email behaviour. The main findings are as follows:

- We observed 2 distinct behaviours in our logs. 1) The user was trying to access the same message (signified by high message click overlap) sometimes by submitting the same, but often a different or similar query. 2) The user was submitting a repeated query but looking for a different message (indicated by low or no message click overlap).
- When a repeat query is submitted there is on average high click overlap. However, this changes with time. The longer the gap between submitting and resubmitting the same query, the less chance they are looking for same message. This is very different to SE behaviour and shows that there is a strong temporal dimension in email search, whereby the information need associated with a query changes with time.
- The lexical similarity of repeat query pairs was a good indicator of click-through overlap. This suggests that similarity combined with the time between queries provide strong clues as to what people want to find.

– People are very important to email search. Lots of queries feature references to people and very often there is high sender overlap in the click through data. Further, most searches heavily feature the people who send the most emails to people (much more than would be expected based on frequency of sending alone).

In the following section we try to reason about what these findings should mean for the design of email clients and email search interfaces.

8 Design implications

Our data show that in different situations, when people resubmit the same or a lexically similar query they could either be 1) looking at a message they have clicked on before or 2) they could be looking for a different message on the same topic or from the same sender. This means that email search systems should behave differently to the web search engines for repeat queries. For SEs it probably makes sense to keep messages that were clicked on the last time the query submitted higher in the rankings because, as SE logs show, there is a very good chance that they are looking for the same page again [15]. However, the temporal aspect of email information needs, revealed by our analyses, suggests that this is probably not the best approach for email. Email search systems need to be smarter in order to predict what the user is looking for. Our findings revealed two clues (time between submitting and re-submitting and the similarity of the repeat query) that could help systems understand what the user is looking for and determine what results are best to show the user. We have shown query overlap is a good indicator of what the user is looking for. Thus, future analyses could investigate how accurately we can predict query overlap based on time gap and query similarity in order to establish if this could help determine which results to show the user.

Another way to deal with the temporal aspect of email information needs is to improve search interfaces to allow the user to indicate how old the message they are looking for is. Current interfaces typically only offer the possibility to sort messages by time. If interfaces could provide an effective means to communicate a temporal aspect to a query this could really help the search process. This could be achieved by some kind of timeline graphical widget where the user could indicate a time period of interest, similar to that suggested for computer files [9].

The finding that the top k senders are searched on comparatively frequently suggests it may be beneficial to provide users with a view on to their emails organised around the top k senders. This could facilitate more effective or quicker access to messages from these senders.

9 Conclusions

In this paper we have presented the first detailed analyses of email search queries. The analyses reported on reveal important insights into how people search for

emails and how this behaviour could be supported by improving ranking functions or search interfaces. We plan to build on the presented work by investigating how message overlap could be predicted and how knowledge of temporal importance can be used to predict which emails have relevance only in the short-term and which are relevant for longer periods of time. Furthermore we want to investigate if certain messages serve as "hubs" or "beacons," assisting in the re-finding of other related emails.

10 Acknowledgements

Thanks to Susan Dumais and Paul Thomas for helpful comments and ideas.

References

1. http://www.facebook.com/press/info.php?statistics Last accessed on 13th September 2011.
2. http://blog.twitter.com/2011/06/200-million-tweets-per-day.html Last accessed on 13th September 2011.
3. E. Adar, J. Teevan, and S. T. Dumais. Large scale analysis of web revisitation patterns. In *Proc SIGCHI*, CHI '08, pages 1197–1206. ACM, 2008.
4. S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D.C. Robbins. Stuff i've seen: a system for personal information retrieval and re-use. In *Proc. ACM SIGIR '03:*, pages 72–79, 2003.
5. D. Elsweiler, M. Baillie, and I. Ruthven. What makes re-finding information diffcult? a study of email re-finding. In *Proc. ECIR 2011*, 2011.
6. D. Elsweiler, M. Harvey, and M. Hacker. Understanding re-finding behavior in naturalistic email interaction logs. In *Proc.ACM SIGIR*, pages 35–44. ACM, 2011.
7. D. Elsweiler, D. E. Losada, J. Toucedo, and R. Fernandez. Seeding simulated queries with user-study data forpersonal search evaluation. In *Proc. ACM SIGIR*, pages 25–34. ACM, 2011.
8. D. Elsweiler and I. Ruthven. Towards task-based personal information management evaluations. In *Proc. ACM SIGIR 2007*, pages 23–30, 2007.
9. E. Freeman and D. Gelernter. Lifestreams: a storage model for personal data. *SIGMOD Record*, 25(1):80–86, 1996.
10. B.J. Jansen and A. Spink. How are we searching the world wide web? a comparison of nine search engine transaction logs. *IPM*, 42:248–263, January 2006.
11. K. Purcell. http://www.pewinternet.org/reports/2011/search-and-email/report.aspx, Aug 2011.
12. S. Radicati. Email statistics report. 2010.
13. M. Sanderson and S.T. Dumais. Examining repetition in user search behavior. In *Proc. ECIR*, ECIR'07, pages 597–604, 2007.
14. C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large alta vista query log. Technical report, SRC, 1998. Technical Note 1998-014.
15. J. Teevan, E. Adar, R. Jones, and M. Potts. Information re-retrieval: Repeat queries in yahoo's logs. In *Proc. SIGIR '07*, 2007.
16. S. K. Tyler and J. Teevan. Large scale query log analysis of re-finding. In *Proc. WSDM '10*, 2010.
17. S. Whittaker and C. Sidner. Email overload: exploring personal information management of email. In *SIGCHI 1996*, pages 276–283, 1996.