Visualising PIM Behaviour with Markov Chains

David Elsweiler
Department Computer Science 8 (AI), University of Erlangen
david.elsweiler@i8.uni-erlangen.de

Martin Hacker
Department Computer Science 8 (AI), University of Erlangen
martin.hacker@i8.uni-erlangen.de

Stefan Mandl
Department Computer Science 8 (AI), University of Erlangen
stefan.mandl@i8.uni-erlangen.de

ABSTRACT

This paper presents our initial efforts at visualising personal information behaviour using Markov Chains. We describe a laboratory-based study of email re-finding and use Markov Chains, created from captured user interactions, as a means of understanding the behaviour exhibited. The models we generate not only provide an excellent overview of how the participants interacted with the experimental interface, but, by forcing the experimenters to ask questions they would not normally ask in order to comprehend the models, they also offer a starting point from which a fuller understanding of the exhibited behaviour can be attained. We illustrate this through examples, discuss the advantages and limitations of the approach and outline how we will expand on the work in future research.

1. INTRODUCTION

A key challenge for PIM researchers lies with evaluation. Few techniques exist to help understand how people use PIM tools and, consequently, very few of the many prototypes that have been designed have actually been evaluated. This lack of tool evaluation has been repeatedly identified as a factor restricting progress in the field, e.g. [2, 7]. By studying how people use PIM tools we can understand what interactive support people need when re-finding, evaluate the effectiveness of existing tools and inform the design of more useful tools for managing and re-finding information.

In this paper we present a novel evaluation method, based on the visual analysis of statistical models derived from interaction logs. Although we are still at an early stage with this work and have only used the method on a small dataset, our findings so far have been very positive and suggest that this approach may be helpful in gaining an understanding of PIM behaviour not possible with existing techniques alone. Here, we explain the approach, illustrate the advantages and limitations using examples collected from a laboratory-based study of email re-finding, and continue to outline our plans to develop the approach in future work.

2. BACKGROUND LITERATURE

The main method for PIM evaluations is to use laboratory-based user studies where users are observed in controlled environments. Such studies can provide an understanding of participants' re-finding strategies, such as the teleporting and orienteering strategies observed in [10] and the discovery that people generally prefer spatial browsing over keyword search [1]. Lab-based user studies can also be used to verify the benefits of particular tools, e.g. [9, 8]. However, there are also limitations to such studies. First, they are performed in artificially created environments with the presence of an experimenter, both of which are likely to impact on the participants' behaviour. Second, when any more than a handful of participants are observed over long time periods it becomes difficult to establish fine grain patterns in behaviour. Further, user studies rely heavily on experimenter observations so the findings are often criticised for being anecdotal and open to subjective bias.

An alternative method to laboratory approaches is to use log-file analysis techniques (LFA) to learn about user behaviour in naturalistic conditions out-with the control of the experimenter [5, 3]. LFA examines the quantitative aspects of user behaviour, including the nature of submitted queries and the properties of items accessed. This is an important technique as it allows the capture of a large quantity of data relating to how users behave with systems without the expense and distracting influence of an observer. The data are also less susceptible to subjective bias. Nevertheless, as the captured data show nothing about what the user is trying to achieve or the tasks that they are performing, it is difficult to make any concrete statements about the reasons for the behaviour depicted in the logs.

In this paper, we propose a method that we believe assists with many of the limitations outlined above and has the potential to formally combine the findings from both kinds of study. By modelling the user's interactions with a system as a statistical process (we use Markov chains), we show that users' behaviour can be visualised in an intuitive way, allowing the experimenter to analyse behaviour retrospectively.

A Markov model is a discrete-time stochastic process which describes the state of a system at successive points in time. Markov Modelling (MM) has been applied in many domains for many purposes, e.g. speech, handwriting and gesture recognition. Modelling techniques have been applied to search behaviour before, particularly in the field of IR. For example, models have been used to improve retrieval algorithms based on prior behaviour e.g [11]. Other models have looked at trying to predict user behaviour such as queries they might apply or results they may click on e.g. [4]. However, using such models to evaluate information behaviour, particularly visually as we do here, is completely novel. We chose to start our work using Markov chains rather than more complicated processes, precisely because they are simple and we believed more suitable for the purpose of visualisation.

3. CONSTRUCTING THE MODELS

We examined the feasibility of the MM approach by using data collected from a user study investigating email re-finding behaviour. 21 participants consisting of a mix of undergraduate and postgraduate students, as well as research and academic staff from the University of Strathclyde each performed 9 re-finding tasks generated according to the method suggested in [7]. Each participant performed 3 tasks on each of the 3 experimental systems. However, for brevity and to simplify the explanation of the MM technique, in this paper we focus purely on one of the experimental systems that will be familiar to all of our readers – a folder-based email client[1]. The experiment included participants who had different quantities of emails (mean = 1938 , sd =2911), used email for different purposes, and who employed different filing strategies.[2]

We created the models by mapping the possible ways that a user could interact with the system to a set of states, using the interaction log data to count the number of times a user moved from one state to another and using the counts to calculate the transition probabilities in the model. We chose states to represent sorting the displayed emails by various attributes, opening a folder, and selecting an email to view its content. We also included a start and two end states – task completed and task abandoned. We present different models in Figures 1, 2, 3 and 4.

One of the advantages we foresaw with this approach was the ability to determine teleporting and orienteering behaviours [10] at a glance. We expected teleporting behaviours to be represented by few states and have high transition probabilities between the states and orienteering behaviours to have many states and low transition probabilities. To test this we derived models based on the interactions for 2 tasks for which we had noted these behaviours during the evaluation (Figures 1 and 2). You can tell that Figure 1 depicts teleporting behaviour, with the user going straight to the messages in the inbox, clicking in total 4 messages, before finding what he needed. This is contrasted with the behaviour in Figure 2, where the user was looking for clues in the messages. He sorted by sender, date and subject and selected 22 emails during the task. This is behaviour indicative of an orienteering strategy.

Figure 3 presents a model generated for all of the tasks performed on the folder-based system. This model provides a good overview of how the participants behaved with the folder-based interface. It shows, for example, that when the participants tended to start their search by sorting, 'sender' was the most frequently used attribute to sort on (~35% of all tasks), while 'subject' was used least often (~11% of all tasks). Considering emails were by default ordered by date, sorting by date was clicked on surprisingly regularly as a first interaction (~21% of all tasks). Although, if users were searching for older mails it makes sense for them to have reversed the order. In ~14% of tasks, the participants chose to open a folder as their first interaction. It seems, however, that participants weren't always sure which folder to search in as ~63% of folder openings were followed by opening another folder. The model also shows that the emails within folders were regularly (at least 13% of the time) sorted by sender. The third major strategy used by participants, after folders and sorting, was to look directly at emails in the inbox. This they did in ~17% of the tasks.

From Figure 3, we see that 'SelectEmail' is a 'hub' state with many in-links, but few out-links, the main out-link being to 'Completed'. 'SelectEmail' also has a very high percentage of looping transitions (~91% of email selects were followed by another). This suggests that after choosing to examine one email, the participants tended to continue to examine other emails until they found what they required or abandoned the task. Combining this observation with the fact that folders and sorting were used mostly at the start of tasks (with the exception of the start state, neither the open folder state nor any of the sorting states had many (if any) in-links), means that the model depicts an overall pattern of behaviour where the participants, firstly, narrowed the search space using sorting, folders or a mixture of both, and then followed this by examining the remaining emails. This is confirmed by following the paths in the model with transitions with the largest percentages. There are, however, two 'reverse' transitions ('SelectEmail' to 'OpenFolder', which represents 10 interactions and 'SelectEmail' to 'SortBySender', which represents 8 interactions). These transitions go against the flow of the narrowing and checking pattern described above, perhaps indicating a change of strategy mid-task. Thus, the main trend in the model is short, direct paths between start and end states, with very little interaction between the states, e.g. the participants did not transition between sorting states. However, the 'reverse transitions' show that the participants didn't always behave in this way. Further examination of the 'reverse transitions' revealed that all but 2 of the 18 interactions came from sequences in which folders had previously been opened.

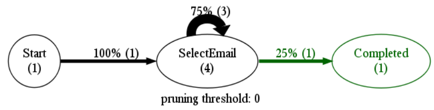

Figure 1: A model for user-id 2, task 3 – teleporting strategy

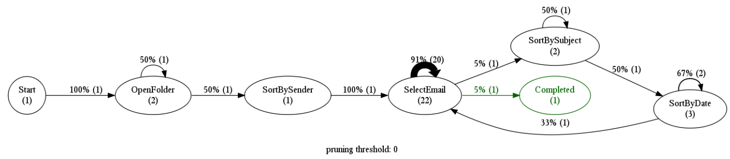

Figure 2: A model for user-id 25, task 1 – orienteering strategy

[1]The interface was based on the Mozilla Thunderbird interface (http://www.mozilla.org)

[2]Due to space restrictions we are only able to provide minimal details regarding the experimental design. However, full details of the tasks and how they were created, properties of the participants and how the tasks and systems were rotated to create a balanced experimental design can be found in [6].

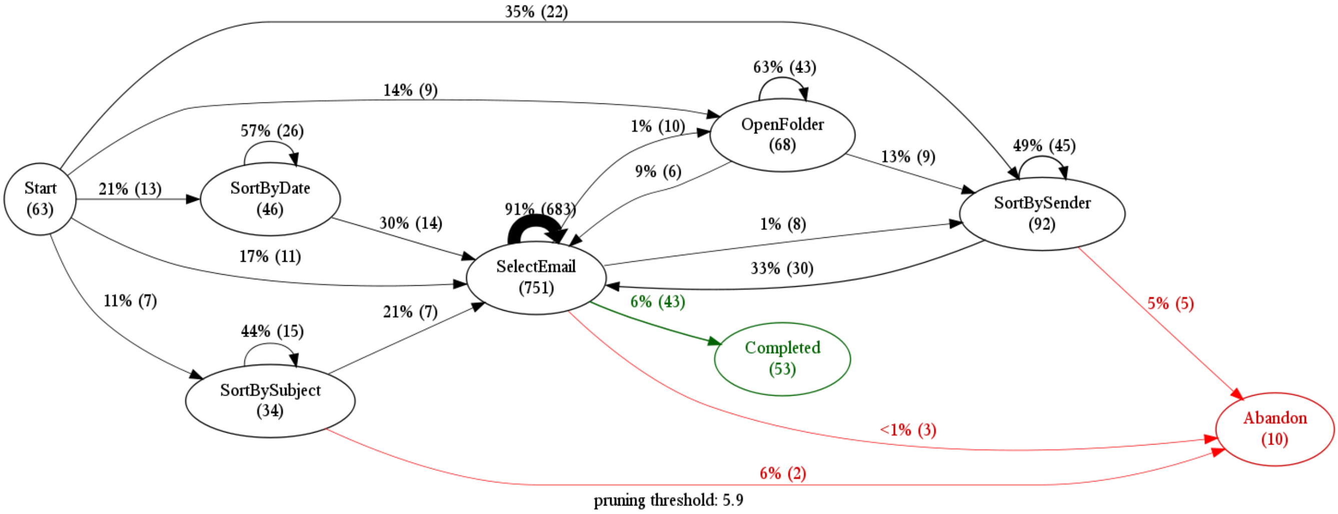

Figure 3: A model generated for all tasks performed on the folder-based system (# tasks = 63). The edges represent the frequency of choosing the target node as next action, the bracketed figures on the edges represent the total number of interactions between these states. The line thickness is also an indicator of interaction frequency with thicker lines representing more interactions. The bracketed figures within the states represent the total number of in-links to that state. We have removed edges representing <5.9 transactions with the exception of the transitions to 'abandon', which we felt helped the reader's understanding of the model and the behaviour it represents.

This suggests that rather than changing strategy to folders or 'SortBySender' mid-search, it seems that the participants needed several attempts to find the correct folder and sometimes required to sort messages in folders by sender to detect this, even after looking at some of the messages in the folder.

The short, direct paths, depicted in Figure 1, are in contrast to Figure 4, which presents a model for the 10 of the 63 tasks that the participants failed to complete. Whereas the first model shows relatively little interaction between the states, the model for failed tasks shows much more interaction between states, with the transitions having lower percentages attached.[3] High interaction with low probabilites suggests longer interaction sequences. This is corroborated by the data. Completed tasks had on average 14.3 interactions, while incomplete tasks had on average 23.3. This is to be expected with incomplete tasks, as when an initial strategy failed the participants would have tended to try other tactics.

The 10 tasks used to generate this second model probably do not provide enough data to establish if the participants utilised different strategies when attempting these failed tasks, i.e. use sorting or folders more often. However, the transitions from the start state seem to be forming a similar pattern to those in Figure 3, suggesting that similar strategies may have been employed. Interestingly, Figure 4 shows that in half of the failed tasks, the participants used a sort by sender in a last attempt to find the required information.

[3]These transitions of course featured in the first model, however because our pruning algorithm removed transactions representing small numbers of interactions these were removed to increase the readability of the model and convey the main trends

4. SUMMARY, FUTURE WORK AND CONCLUSIONS

The examples we have provided demonstrate how visualising users' interactions as Markov chains can allow experimenters to understand how users behave with a system by offering the opportunity to analyse behaviour visually. We were able to identify several aspects of behaviour, including how sorting and folders were used and recognise teleporting and orienteering behaviours. While visualising interactions in this way allows complicated datasets to be understood, one limitation is that this understanding cannot be gained simply by glancing at the models – they need to be studied in depth and this process requires no little creativity on the part of the experimenter. Nevertheless, the positive aspect is that the process of analysing the models forces the experimenter to ask questions he would not otherwise ask, leading to a better overall understanding of what is going on. A good example of this was examining the 'reverse transitions' as described above. Analysing the models can also lead to the generation of new research questions e.g. why was 'SortBySubject' a common last resort in failed tasks?

We also showed that it is possible to use the Markov chains as a means to visually compare behaviour in different situations. The example we provided compared all tasks with incomplete tasks, but the approach could be used, for example, to compare the behaviour of different types of user (e.g. experienced vs. novice users, filers vs. pilers, older vs. younger participants etc.), behaviour for different types of task (e.g. looking for older or newer information), or for different types of systems (e.g. browse-based vs. search-based). We are currently building these models and looking at ways in which interactions can be abstracted so that different systems can be compared. We are also exploring methods of mathematically comparing models that can be used to au-

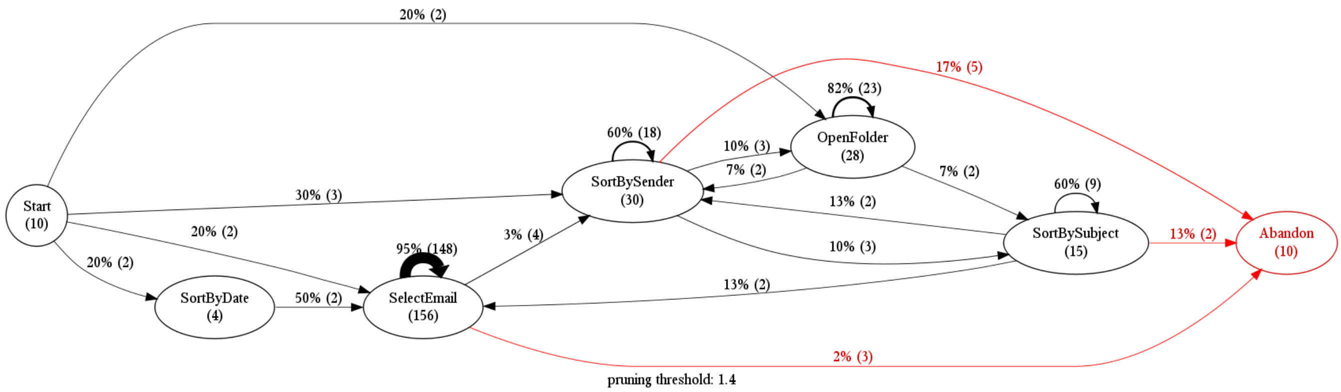

Figure 4: A model generated from all failed tasks on the folder-based system (# tasks = 10). To ease readability we have removed edges representing <1.4 transitions

tomatically detect the kind of features we observed when visually analysing the models (e.g. hub states, high or low interaction between states, probable paths etc.). This would help researchers identify behavioural changes and corroborate any observations made, as well as lessen the reliance on the experimenter's creativity when analysing the models.

We must mention some dangers with analysing data in this way. Pruning, for example, makes it easier to analyse the models and spot patterns, but it can be misleading. It is extremely important to verify hypotheses generated from the pruned model on the original un-pruned version. Another danger is that looping transitions can lead to misunderstandings. For example, although the transition from 'OpenFolder' to 'SortBySender' in Figure 3 has an associated probability of ~13%, actually, if you discount the looping transitions, this percentage would be closer to 40%. In other words, when the participants were satisfied that they had found the correct folder, they tended to sort by sender very often. This is not very clear from the model as it is somewhat disguised by the looping transition. It is important that experimenters are aware of such properties.

A limitation of Markov chains as we have presented them here is that they have no means to model temporal information, which has been shown to be useful in PIM and search behaviour [8, 4]. We plan to extend our work using different kinds of models to investigate the usefulness of temporal information in this context. However, first we plan to exhaust the potential with simple chains.

Although we have demonstrated the approach using data collected from a laboratory-based study, we believe it will offer even greater potential in the context of naturalistic studies. Naturalistic studies provide far larger quantities of data to work with, which would offer greater scope for patterns to emerge in the data and for mathematical analyses. We are in the process of planning a large scale log-based, naturalistic study of email behaviour and aim to use the MM approach to help analyse the data. A further benefit of the MM approach in this context is that it may allow the findings of the naturalistic study to be formally triangulated with those derived from lab-based studies. If we can find ways to mathematically compare sequences of interactions to models constructed from behaviour observed by experimenters it would go a long way to overcoming many of the limitations described in Section 1.

5. REFERENCES

[1] D. K. Barreau and B. Nardi, *Finding and reminding: File organization from the desktop*, ACM SIGCHI Bulletin **27** (1995), no. 3, 39–43.

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

[3] E. Cutrell, D. Robbins, S. Dumais, and R. Sarin, *Fast, flexible filtering with phlat*, Proc SIGCHI, 2006, pp. 261–270.

[4] D. Downey, S. T. Dumais, and E. Horvitz, *Models of searching and browsing: Languages, studies, and application*, IJCAI, 2007, pp. 2740–2747.

[5] 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*, Proc. SIGIR, 2003, pp. 72–79.

[6] D. Elsweiler, M. Baillie, and I. Ruthven, *Exploring memory in email refinding*, ACM Trans. Inf. Syst. **26** (2008), no. 4, 1–36.

[7] D. Elsweiler and I. Ruthven, *Towards task-based personal information management evaluations*, Proc. SIGIR, ACM Press, 2007, pp. 23–30.

[8] M. Ringel, E. Cutrell, S. Dumais, and E. Horvitz, *Milestones in time: The value of landmarks in retrieving information from personal stores.*, Proc. INTERACT, 2003, pp. 184–191.

[9] G. Robertson, M. Czerwinski, K. Larson, D. C. Robbins, D. Thiel, and M. van Dantzich, *Data mountain: using spatial memory for document management*, Proc. UIST, 1998, pp. 153–162.

[10] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger, *The perfect search engine is not enough: a study of orienteering behavior in directed search*, Proc. SIGCHI, 2004, pp. 415–422.

[11] J. Teevan, S. T. Dumais, and E. Horvitz., *Personalizing search via automated analysis of interests and activities*, Proc. of SIGIR, 2005.