Pseudo-Desktop Collections and PIM: The Missing Link

Daniel Gonçalves
INESC-ID / IST / TULisbon
Av. Rovisco Pais, 1
1049-001 Lisbon, Portugal
+351 213100248
daniel.goncalves@inesc-id.pt

ABSTRACT

Personal Information Management has for a long time faced a serious problem: validating its results. By dealing with personal information, it is hard to collect performance and quality metrics, and to have a ground case against which possible solutions might be compared. Some efforts have been made to create canonical sets of data that might be used as the basis for such tests. We discuss to what extent are those data sets adequate for PIM, and how they might be improved. We argue that they capture only a limited part of the information in play in real scenarios, and while useful have a restricted applicability. Much meaning is provided by the users themselves, making it hard for information sets not annotated with such meta-data to suffice.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: [Information Search and Retrieval]

General Terms

Design, Standardization, Theory

Keywords

Personal Information Management, Reusable test collections, pseudo-desktop, desktop search, information retrieval

1. INTRODUCTION

The area of Personal Information Management (PIM) is concerned with the study of how people manage their information, from organization to retrieval. There have been many attempted solutions to those problems, and all have ultimately faced the same problems: Adequacy and Evaluation.

Regarding Adequacy, there is a wide range of ways in which individual users manage their information. Thus, while it is possible to test a new system for correctness with a custom-made or well known set of information (often the researcher's own), the doubt remains about whether the solution will work in the general case. Tests done with limited numbers of datasets are anecdotal at best, and there is a risk of over-specialization. Thus, extensive user tests must be performed, leading to the Evaluation problem.

When performing user studies for interactive systems, it is customary to ask users to perform a set of pre-determined tasks. Usability metrics such as task completion times and number of errors are then measured and used as a basis for discussion. In other areas, such as information retrieval, retrieval methods are applied to known (and often pre-classified) datasets. This enables the calculation of measures such as precision and recall. When evaluating PIM solutions, neither is possible. Since we're trying to evaluate solutions that deal with the users' own information, even if we ask the users to perform a same task, *the tasks they end up doing are not the same*. Specific tasks are dangerous. We cannot, for instance, ask all users to find "a document they wrote about dogs". Many won't have such a document. Only more general tasks are possible, such as asking users to find "the last document they wrote and sent to someone", but then the steps that might need to be done might differ wildly from user to user. One might have done so yesterday, the other a couple of months ago, in different settings, for different purposes, etc. Also, it is impossible for researchers to know the users' information and thus know for sure if a particular task has succeeded (was a document not found because of system failure or because it wasn't there to be found in the first place?).

All this makes the validation of PIM systems hard, and begets the creation or definition of a meaningful, representative, set of personal information that can be used as the canonical basis for testing. Such a set would solve the Adequacy problem, and alleviate the Evaluation problem, by allowing researchers to know the data beforehand. While replacing user studies is impossible, such sets might suffice to find meaningful preliminary results and as a way to compare solutions. This was attempted by Kim and Croft [3]. The authors produced three sets of pseudo-desktop data and associated queries. Their goal was to provide information that might be used to evaluate desktop search systems.

2. REPRESENTATIVENESS

The test collections described in [3] contain information divided into five categories: HTML pages, Emails, Word, PDF and Powerpoint files. All with the exception of emails contain around 1,000 items (emails are an order of magnitude more). This is not necessarily representative of a real personal information collection. Previous studies [2] have found a different distribution for the different item types. Another factor that could be taken into consideration is the relatively high variability in personal information collections. Three classes of users were identified, and it would be advantageous if the three pseudo-desktop collections reflected those types. Another important omission are multimedia files. Images and video have a growing importance in the users' lives. The sets should reflect this. Finally, there is organization information missing, folder hierarchy being the most important absence. This information reflects how users organize their information and are important to understand their real needs.

3. SUITABILITY FOR PIM RESEARCH

It is out opinion that, in their present form, the sets might be useful for specific retrieval-related solutions, but are in general unsuited for use with PIM tools. Our objections relate to three related key aspects: Lack of Autobiographic Information, Lack of Meaning, and Lack of Ground Truth.

3.1 Lack of Autobiographic Information

Personal information doesn't exist in a void. It has been previously handled by users, for a reason, in a context. This context goes beyond the computer, into the users' personal and professional lives. It is related to an extended set of autobiographic information, implicitly present in the users' minds. A user might know a document was written around his son's birthday, or during a relative's illness. This information is not part of the documents, but important to users and will determine how they and other personal information are remembered. By automatically generating pseudo-desktop collections, all autobiographic information is missing. It will be possible to use the collections to test techniques for which only the data in the documents is relevant, but not those for which the context is important. Autobiographic information often determines how and why tasks are performed. Furthermore, it is possible to design solutions where it plays a central role, by allowing users to manage their information in *personally relevant ways*.

3.2 Lack of Meaning

Consider an email message. Everyone might look at its sender, `johndoe@somewhere.com` and know that someone at that address is the recipient of the message. From the point of view of the user that sent it, the recipient isn't that email address, but rather (for instance) John Smith, a person with a shared context, other times referred as "Johnny" or even "boss". There is more meaning than can be gleaned from the email message itself (although it might to some extent be inferred from the entire data set). Also, when discussing "the project" in a message to that person the user might know that, in that context, "the project" is actually "Project Foo", on which both work. When designing a retrieval tool, it might make sense for searches for emails to "the boss" to return those to johndoe and "Johnny", etc. Without the users' knowledge, it will be very hard for a system to know they all represent the same person. Having such meaning would also allow us to test how solutions address the well-known Fragmentation Problem [1].

3.3 Lack of Ground Truth

It would be interesting to have data classified according to personal criteria. In traditional retrieval solutions, the set of documents is often manually classified to allow measures such as precision and recall to be computed. This also allows task success to be evaluated. In the context of PIM things are more complex. If a user requests documents about "Subject X", what should be returned? Most likely, not only those that actually contain the words "Subject X", but also others, related to that subject in some way (not to mention multimedia files for which there is no textual information at all). Paraphrases, synonyms, related people and subjects, might all be needed to take into consideration. Again, the user is often the only one that can provide this information, not only complex, but also of a subjective nature. The actual results that would satisfy the user might even change according to the context *at retrieval time*. Having this kind of ground truth would be necessary to evaluate PIM solutions.

4. WILL A SOLUTION EVER EXIST?

The Lack of Autobiographic Information looks at the wider context in which the information is used, and is extrinsic to the data set. The Lack of Meaning reflects the need to have an overall integrated view of all the information. The Lack of Ground Truth is related to how users view their data.

These problems point to the way to create information sets useful and reusable for the evaluation of PIM tools. First and foremost, real information from real users must be collected. An updated study to identify archetypical user classes must be performed, and a different user selected for each class. The collected information must include a wealth of data sources (files, email, calendar, contacts, etc). There are major privacy issues to be addressed. Most can be solved by anonymizing the data, consistently exchanging real names and addresses by simulated ones. A deeper level of anonymization might be necessary, handling project names, places and other sensitive information. This is the simplest part of the creation of the information set.

The users' cooperation would be necessary for the next steps: annotating the information with subjective metadata. The users would need to use a special purpose tool to enter autobiographic information. Also, they would be asked to annotate the documents themselves (and other information), minimizing the Lack of Meaning problem. Finally, they would be asked to classify their documents according to high-level tasks and subjects, addressing the Ground Truth problem (using tags instead of hierarchies, as the same information item might have different uses and meanings). Part of this might be done automatically. For instance, if two email messages are sent to `"John Smith <jsmith@gmail.com>"` and `"John Smith <johns@hotmail.com>"`, the system can make the educated guess that both are the same person. But still this would need to be checked and complemented by the user. The process would be iterative, to fine tune the result. Also, by monitoring the users' everyday use of their information, a set of representative tasks and queries should be collected.

It would be a labor intensive process, but result in information sets that can be understood even in the absence of the user, and used in a rich set of situations where personal information and its surrounding context are relevant.

5. CONCLUSIONS

The creation of pseudo-desktop collections is a worthy goal. Such sets might be very important in providing a testbed for repeatable, comparable experiments, and greatly facilitate the validation of PIM tools. Current versions lack key elements related to the users and the context in which the information is used, which will have to be included for the sets to be of use in a broader context.

6. ACKNOWLEDGMENTS

This work was supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds and FCT project EDUCARE (PTDC/EIA-EIA/110058/2009).

7. REFERENCES

[1] Bergman, O., Beyth-Marom, R. and Nachmias, R.. The project fragmentation problem in personal information management. In Proc. CHI '06: pp. 271–274. ACM Press, New York, NY, USA, 2006. ISBN 1-59593-372-7.

[2] Gonçalves, D., Jorge, J., An Empirical Study of Personal Document Spaces.. In DSV-IS'03. LNCS v2844. pp. 47-60. Springer-Verlag, 6-9 June 2003, Funchal, Portugal

[3] Kim, J. and Croft W. B. (2009) Retrieval experiments using pseudo-desktop collections. In CIKM'09, pp1297–1306. ACM.