Managed Forgetting, Data Condensation & Preservation in Application

Christian Jilek [a]
christian.jilek@dfki.de

Heiko Maus [a]
heiko.maus@dfki.de

Sven Schwarz [a]
sven.schwarz@dfki.de

Andreas Dengel [ab]
andreas.dengel@dfki.de

[a] Knowledge Management Department, German Research Center for Artificial Intelligence (DFKI) GmbH, Trippstadter Straße 122, 67663 Kaiserslautern, Germany

[b] Knowledge-Based Systems Group, Department of Computer Science, University of Kaiserslautern, P.O. Box 3049, 67653 Kaiserslautern, Germany

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
Ubicomp/ISWC'16 Adjunct , September 12-16, 2016, Heidelberg, Germany
© 2016 ACM. ISBN 978-1-4503-4462-3/16/09...$15.00
DOI: http://dx.doi.org/10.1145/2968219.2968567

Abstract

With an increasing amount of available sensors lifelogging produces more and more data. Thus, realizing necessary condensation and forgetting processes becomes a challenge. In the last three years we investigated Managed Forgetting, Synergetic Preservation and Contextualized Remembering in the so-called *ForgetIT* project. Using the Semantic Desktop as an ecosystem we have already applied these approaches to personal information management successfully. With these experiences at hand we think that lifelogging could also benefit from these solutions. On the other hand, achievements and findings of the lifelogging community can help us in realizing one of our newest visions. In this paper we will provide more details of our data condensation, preservation and managed forgetting solutions and show how lifelogging could benefit from them. Additionally, we sketch our newest application scenario of a context-focused work environment that will make use of lifelogging technologies.

Author Keywords

semantic desktop, diary, data condensation, preservation, managed forgetting

ACM Classification Keywords

H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

Motivation

In their comprehensive summary [3], Gurrin et al. name several challenges and issues lifelogging still faces today. Beside challenges in the areas of capturing and accessing data, ownership or privacy concerns they also discuss the aspect of forgetting. Some researchers see lifelogging as "an antithesis of forgetting, while others aim at modelling the human experience of forgetting in surrogate memories" [3]. In our recently completed EU project called *ForgetIT*[1], we investigated Managed Forgetting, Synergetic Preservation and Contextualized Remembering. Roughly speaking, we conceived forgetting as a means to focus on important things while neglecting irrelevant details. On the one hand, certain information items are of such importance that they need to be preserved even longer than a person's own lifetime. On the other hand, a lot of information is only relevant for a small amount of time and can hence be condensed or even discarded. Beside preserving certain items for the future, this also aims at reducing daily information overload. Typical IT systems usually do not have such tidy up mechanisms, so a user's computing devices become more and more cluttered with information over time. *ForgetIT* considered the aforementioned approaches in the context of personal information management (PIM) aiming at supporting knowledge workers while being embedded in their daily work. Instead of capturing everything that is possibly available, the idea is to rather have a targeted selection of information sources and sensors according to the current and (planned) future situations. Moreover, we think that the solutions found in *ForgetIT*, which will be presented in more detail in the main part of this paper, could also be very beneficial for the lifelogging community. In particular, *ForgetIT* technology can be used to collect (and filter) all information that determines/influences the professional as well as private life of a person (e.g. documents, calendar events, tasks, etc.) On the other hand, we also see aspects of lifelogging which are particularly useful for our future work or PIM in general. Using lifelogging technologies we could get even more insights into the life of the users, especially their different activities in order to consolidate our information model.

The rest of this paper is structured as follows: First, we will give a short introduction to the *Semantic Desktop*, which is the ecosystem our approach is based on. The main section contains more details about the solutions we found and developed in *ForgetIT*. Last, we conclude this paper and show some aspects our research will be focused on in the near future.

Semantic Desktop & PIMO

Like stated before, we base our approach on the concept of the *Semantic Desktop* (SD) [1, 8]. Its core idea is bringing *Semantic Web* technologies to the user's desktop[2]. Since these standards based on ontologies allow representing and organizing data across application borders [2], it is therefore possible to explicitly express (major parts of) a user's personal mental model and make use of it in all their applications – or at least in those that integrate into their personal knowledge space. One of the SD cornerstones is the personal information model (or PIMO for short) [9] which serves as the basis for knowledge representation and provides a common vocabulary across different applications. It consists of *concepts* (called "things" such as specific topics, projects, persons, tasks, ...), *associations* between them (persons are *member of* projects, a task *has topic* SD, ...), and finally, *associated resources* (documents,

[1] http://www.forgetit-project.eu/

[2] The term *Semantic Desktop* dates back to a time when desktop computers were the most prominent computing devices. Today, this term also comprises smartphones, tablet computers, etc.

e-mails, web pages, pictures, ...) [5]. The most recent SD implementation is in the form of a cloud-based service providing a service API based on JSON RPC that uses the PIMO schema with its classes and properties and intended semantics, relies on URIs to identify things and resources, and most importantly, defines a set of methods to access and manipulate the PIMO [5].

The SD architecture integrates various sources for real-life events (e.g. eye tracking data [11]). Lifelogging technology provides a successful means to capture a continuous stream of detailed information about the user's environment. Integrated into the SD, we enrich the PIMO of the user and increase the level of details and authenticity. Additionally, the (lifelogging) data is put into semantic context, thus enabling semantic indexing, condensation and forgetting.

In the *ForgetIT* project we took the SD as basis and realized approaches like Managed Forgetting, Data Condensation and Preservation, which are discussed in more detail in the next section.

Managed Forgetting, Data Condensation & Preservation in the Semantic Desktop

Managed Forgetting in the SD In the SD Managed Forgetting is used in several ways. Based on user interaction data and the semantic graph, we calculate the so-called *memory buoyancy*, which represents an information item's short-term value. The main idea is that an item's importance rises and falls according to observed user actions, calendar events, incoming emails, documents relating to similar topics, etc. As a consequence, when users browse their PIMO, only the currently most relevant items are shown. Other items are initially hidden and are only shown on explicit user request or if their estimated importance rises again.

Another aspect of Managed Forgetting deals with synchronizing files to mobile devices. First of all, all items are stored on the server, but in order to allow quick and offline access, resources with high buoyancy are cached on portable devices. Depending on the devices' capabilities (especially regarding disk space), the caching is more or less restricted: Smartphones and tablets only keep resources with very high buoyancy whereas laptops also keep resources of medium to high importance. Technically this means that for different devices we use different buoyancy thresholds for synchronization. Additionally, as the SD keeps different data structures for resources and their metadata, the metadata (like author, topics, etc.) are treated differently to the resources (they use different buoyancy thresholds), that is, they are kept much longer on the devices than the resources. So, for example, if the user has not worked on a document for a month, the document will be removed from the document cache on the tablet. However, the metadata is still present and, hence, the document can be found in a search request on a train even when being offline. In that case, to open the document the user would have to either get online again or use his laptop (which has a higher buoyancy threshold). To complete the example: Also metadata can get large over many years. Thus, the devices will not keep *all* metadata of the whole PIMO. For metadata, the caching simply uses a much lower buoyancy threshold than for resources. Surely, for all automatisms one can create worst case counter examples. This is not different for the SD, but it is also true for any IT system with automatisms, including manual, human secretary work.

A third aspect is about data condensation. Instead of keeping all data of an event, for example, only the most important details are preserved forming a kind of memory landmark. Forgetting means hiding in most cases, however we investigate scenarios where we actually delete resources with long-term low buoyancy. We want to contribute to the lifelogging community by proposing to use similar forgetting techniques to cope with the endless stream of recorded data: by condensing or forgetting irrelevant portions of data, the lifelogging stays scalable. More details about this are given in one of the next subsections about *PIMO Diary*.

Preservation in the SD As a counterpart to the aforementioned memory buoyancy we also use the so-called *preservation value* [7], which represents an information item's long-term value, i.e. whether an item is worth preserving for the future. Similar to the approach given in [10], we implemented a module constantly assessing all of a PIMO's resources according to their possible preservation worthiness. We started with the application scenario of personal photo collections [12] and later added some more general rules/heuristics for all resources. When calculating a resource's preservation value six dimensions are taken into account:

- **investment**: The more effort a user invested on a certain resource (e.g. number of annotations/tags made, length of a comment or other usage statistics), the higher its probability of being preserved should be.
- **gravity**: If an item is (semantically) closer to an important event, person, project, etc., its preservation probability should be higher.
- **social graph**: If an important person is shown on a photo it is preserved more likely.
- **popularity**: The higher an item is rated by the user the more likely it is preserved.
- **coverage**: For example, at least one photo of each of a user's photo collections should be preserved.
- **quality**: High quality photos should more likely be preserved.

In our current use case, we restricted quality and coverage to photos and photo collections, respectively, other measurements are possible, though. These are only some exemplary aspects associated with the different dimensions. Users may select one of four pre-defined preservation profiles. Based on work by Wolters et al. [12], we basically distinguish between so-called *curators* and *filers*. The former take much care in curating their data, for example by adding keywords, writing comments, use face detection tools, etc., whereas the latter more or less just rely on folder names and structures. For more details and sub-variants, please see the original paper or our web documentation[3]. After choosing a profile, the different preservation settings (i.e. options for each of the six dimensions) are adjusted accordingly. Figure 1 shows an example for one of the four profiles. Users may also fine tune them by manually enabling or disabling the different preservation rules/heuristics.

Data Condensation with PIMO Diary Using the SD regularly leads to a PIMO enriched with lots of semantically annotated information, e.g., documents, web pages, emails, photos, calendar events, etc. Sorting, mentally connecting and abstracting from parts of these things in order to remember what actually happened in a given period of time is typically a difficult and time-consuming task. The same

[3] https://pimo.opendfki.de/wp9-pilot/preservation_sd.html

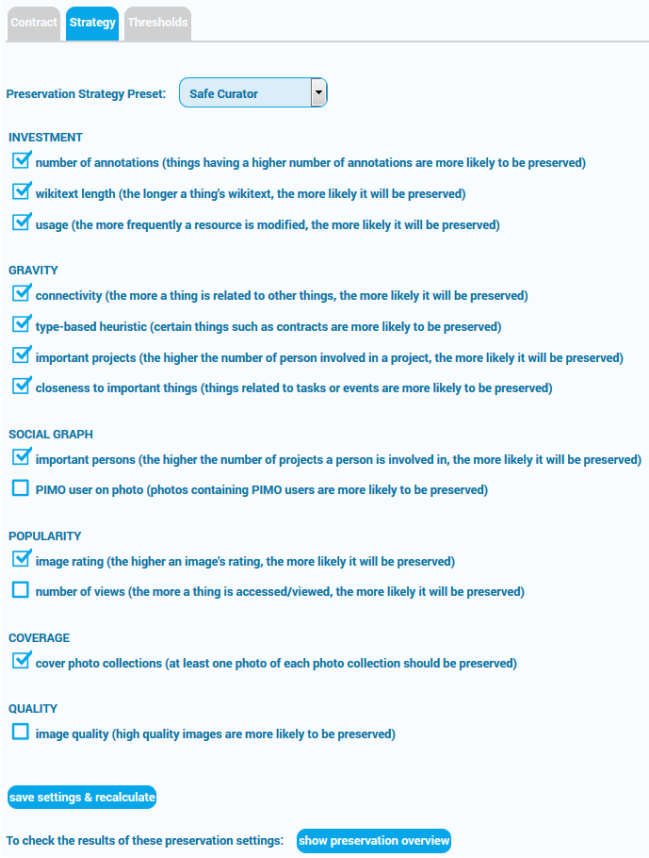

Figure 1: The default settings for the preservation profile of a *Safe Curator*. We assume that *investment*, *gravity* and *coverage* are most relevant for this profile, whereas *quality* is not checked.

is true when trying to find certain events in a large collection of photos and/or videos obtained by users' lifelogging devices.

PIMO Diary realizes contextual remembering by enabling a user to generate a personal (or group) diary based on these information items from the PIMO.

To allow for contextual remembering and at the same time to prevent the diary from being a confusing, large, sequential collection of material, we need to identify semantic relationships among possibly several thousands of individual information items and create suitable abstractions from them. When looking back on the last decade, for example, users should not be overwhelmed with a view showing plenty of individual events, but compact statements like project names or life situations like *school years*, *studies*, *marriage* or the name of a place where a vacation or longer stay abroad has been spent. The user literally *zooms out* of an overwhelming mass of details. If desired, these abstractions can easily be resolved by selecting a sub-period of time for concretization (*zooming in*), e.g., a year of a decade or a month of a year. These concretizations can be performed until the actual basic material (i.e. documents, emails, etc.) is reached.

The system applies a combination of merging and filtering by clustering related or very similar things to diary entries and evaluating their importance for the user. The former aspect fosters a high diversity within the diary, making it interesting and fun to read, whereas the latter aspect is a necessity induced by the fact that the number of diary entries to be generated is usually limited.

Like depicted in Figure 2, a typical diary entry consists of a date or time interval (a), a generated headline (b), and the most prominent things (c) and keywords (d) gathered

Figure 3: The overall diary context should provide a quick overview of those things of a user's life (reflected by their PIMO) that concerned them the most in a given time period.

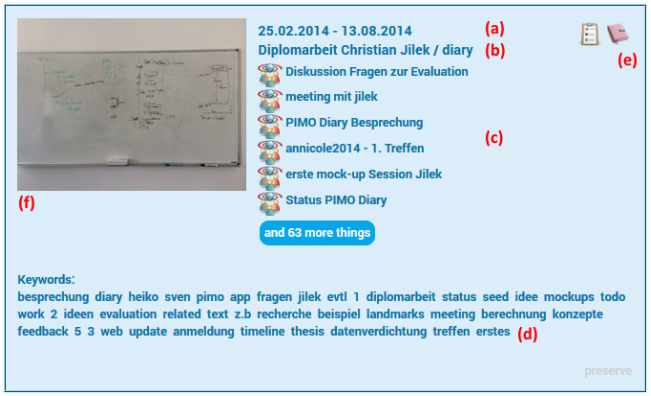

Figure 2: A diary entry consisting of a date or time interval (a), a generated headline (b), the most prominent things (c) and keywords (d), the most prominent annotations (e) and possibly a photo or image (f) is associated with the entry.

from all information items that were clustered to form this entry. Being a cluster of semantically similar items and/or items that have been worked on at the same time, this entry forms already some kind of contextual closure around the contained items. This example is an entry about writing a master thesis and consists of 69 information items. Most of them are calendar events and notes (the screenshot shows the six most prominent ones). On an entry's right-hand side there are its most prominent annotations (e) revealing more of its contextual background. In the example, this is a task (e.g. created in the user's calendar) which was about writing the thesis and a diary which was the thesis' main topic. Additionally, if a photo or image is associated with the entry, it is displayed on its left-hand side (f). In our example, the entry is associated with a photo showing a whiteboard with results of a brainstorming meeting.

Figure 3 shows a diary's overall context, which should provide a quick overview of those things of a user's life (reflected by their PIMO) that concerned them the most in a given time period. In the case of our master thesis example (German Diplomarbeit), we see that these were topics of the thesis (diary, PIMO, data mining, etc.), involved persons and organizations.

Users also have the possibility to incorporate shared data of their family, friends or colleagues - represented by a *group information model (GIMO)* - into their own personal diary turning it into a *group diary*. As a consequence, a friend's shared photo collection appears as a separate entry in a user's own diary or some of their own entries are complemented by additional information items coming from other people's PIMOs, for example.

Our web documentation [6] shows an example of a diary from a productively used PIMO and possible interactions with it. Further details about PIMO Diary were published in [4].

As mentioned above, the diary presents more or less a temporarily ordered set of contexts. In the diary application these contexts were computed automatically. There are cases, however, where users are explicitly confronted with contexts and are attentionally aware of each individual context. We are currently realizing a work environment which allows the user to explicitly focus on one context at a time, thus, showing all context-relevant resources prominently while hiding others. This will be explained in the next section. In that setting, the explicit handling of context will lead to more precise contextual computation of buoyancy and preservation values and, hence, lead to a more direct contextual condensation and forgetting.

Towards a Context-Focused Work Environment

In the future, we will be increasing efforts towards more context-focused work. Let us consider the IT support scenario. One of our recent projects called *supSpaces*[4] tries to improve the situation of support workers that have to cope with occurring issues (called incidents or problems and mostly manifested as "tickets"). Their task is to solve these issues as fast as possible. To do so, they need to (1) focus on the issue and (2) take into account issue-specific information. This task requires special user interface and automatic analysis components.

We provide a context-sensitive work environment (the current context is the issue), which excels at displaying context-relevant information while hiding context-irrelevant data, thus, helping the worker to focus and aiming at removing distraction. The context-specific presentation and filtering can be done very well with SD technology as the SD already provides semantic relationships and automatic concept classification which will be used, for instance, as similarity measures for contexts. Other semantic technologies like semantic filtering, clustering, faceted search, etc. will help filtering for context-relevant information only.

In the planned context-sensitive work environment, context-relevant material is not just explicitly modelled or annotated information elements; instead, observed historical events, actions, and sensory data plays an important role, too. Using similar technology as in lifelogging, the IT support domain also has historical sensoric data. Therefore, one task of the envisioned context-focused work environment is to automatically analyze and categorize these kinds of data and organize them unsupervised "into" the contexts. The data must be analyzed in an online fashion as this data is typically received continuously in form of streams – think of an alarm bus or other sensory buses. Coping with such a large amount of data could easily lead to a confusing mess of information. To prevent this we apply our previously introduced forgetting and data condensation technology. Using these mechanisms we are able to keep the system tidied up, important information is easily found, rather unimportant data is condensed, hidden by default or forgotten.

[4] *supSpaces*: semantic support knowledge spaces for IT support, http://www.supspaces.de/

Conclusion

In summary, we presented several approaches investigated in the *ForgetIT* project, that have already been successfully applied in personal information management and could also be beneficial for the lifelogging domain. First of all, the Semantic Desktop approach with the Personal Information Model semantically representing the user's mental model will contribute to sense making of lifelogging data and applications as well as benefit from more detailed sensory input. Particularly, the automatic computation of values that guide automatic condensation and forgetting should play a major role in the handling of massive lifelogging data. Additionally, we recommend that lifelogging should also focus on a more explicit handling of contexts. Massive data can be handled much better when clustered into contexts. On the other hand, to realize our newly envisioned context-focused work environment we will benefit from solutions and experiences achieved by the lifelogging community.

Acknowledgements

This work has been partly funded by the EU in the IP ForgetIT (GA 600826) and by the German Federal Ministry for Education and Research in the project supSpaces (grant no. 01IS15013B).

REFERENCES

1. A. Dengel. 2007. Knowledge technologies for the social semantic desktop. In *Knowledge Science, Engineering and Management*. Springer, 2–9.
2. A. Dengel (Ed.). 2011. *Semantische Technologien: Grundlagen. Konzepte. Anwendungen.* (1. Aufl. ed.). Spektrum Akademischer Verlag, Heidelberg.
3. C. Gurrin, A. F. Smeaton, and A. R. Doherty. 2014. Lifelogging: Personal big data. *Foundations and trends in information retrieval* 8, 1 (2014), 1–125.
4. C. Jilek, H. Maus, S. Schwarz, and A. Dengel. 2015. Diary Generation from Personal Information Models to Support Contextual Remembering and Reminiscence. In *Workshop on Human Memory-Inspired Multimedia Organization and Preservation (HMMP). IEEE Int. Conf. on Multimedia and Expo (ICME), Torino, Italy*. IEEE, 1–6. http://dx.doi.org/10.1109/ICMEW.2015.7169753
5. H. Maus, S. Schwarz, and A. Dengel. 2013. Weaving Personal Knowledge Spaces into Office Applications. In *Integration of Practice-Oriented Knowledge Technology: Trends and Prospectives*, M. Fathi (Ed.). Springer, 71–82.
6. H. Maus, S. Schwarz, C. Jilek, and B. Eldesouky. 2015. ForgetIT Personal Preservation Pilot. Web documentation. (2015). https://pimo.opendfki.de/wp9-pilot/
7. C. Niederee, N. Kanhabua, F. Gallo, and R. H. Logie. 2015. Forgetful Digital Memory: Towards Brain-Inspired Long-Term Data and Information Management. *ACM SIGMOD Record* 44, 2 (2015), 41–46.
8. L. Sauermann, A. Bernardi, and A. Dengel. 2005. Overview and Outlook on the Semantic Desktop.. In *Proc. of the 1st Workshop on The Semantic Desktop at ISWC*.
9. L. Sauermann, L. van Elst, and A. Dengel. 2007. PIMO – a Framework for Representing Personal Information Models. *Proc. of I-Semantics* 7 (2007), 270–277.
10. T. Tran, S. Schwarz, C. Niederée, H. Maus, and N. Kanhabua. 2016. The Forgotten Needle in My Collections: Task-Aware Ranking of Documents in Semantic Information Space. In *Proc. of the 1st ACM SIGIR Conf. on Human Information Interaction and Retrieval (CHIIR-16), March 13-17, Chapel Hill,North Carolina, USA*. ACM, ACM Press.
11. L. van Elst, M. Kiesel, S. Schwarz, G. Buscher, A. Lauer, and A. Dengel. 2008. Contextualized Knowledge Acquisition in a Personal Semantic Wiki. In *Knowledge Engineering: Practice and Patterns. Proc. of the 16th Int. Conf., EKAW 2008, Acitrezza, Italy, September 29 - October 2, 2008. (LNCS)*, A. Gangemini and J. Euzenat (Eds.), Vol. 5268. Springer Berlin / Heidelberg, 172–187. DOI: http://dx.doi.org/10.1007/978-3-540-87696-0
12. M. K. Wolters, E. Niven, M. Runardotter, F. Gallo, H. Maus, and R. H. Logie. 2015. Personal Photo Preservation for the Smartphone Generation. In *Proc. of the 33rd Annual ACM Conf. Extended Abstracts on Human Factors in Computing Systems (CHI-15)*. ACM, 1549–1554. http://doi.acm.org/10.1145/2702613.2732793