Data Recovery from Distributed Personal Repositories

Rudolf Mayer[1], Robert Neumayer[2], and Andreas Rauber[1]

[1] Institute of Software Technology and Interactive Systems,
Vienna University of Technology, Austria
[2] Department of Computer and Information Science,
Norwegian University of Science and Technology, Trondheim, Norway

Abstract. We present an approach to personal disaster recovery, e.g. after a hard-disk crash, based not on an explicitly ex-ante defined recovery plan with a rigid backup regime, but rather on naturally accumulated and distributed sources of personal data, such as e-mails and their attachments. We aim to restore as much data as possible, and to provide means to organise it in a meaningful folder structure. Employing information retrieval techniques, we semi-automatically establish the context of and relations between the data objects along several different dimensions, to identify relations and groups. Different views at multiple levels of granularity then allow an interactive organisation into folders.

1 Introduction

Disaster recovery is commonly associated with database systems, professional archives, or large businesses. Even though private users or small and medium enterprises are often lacking either financial means or technical skills needed to apply large-scale rescue and backup strategies, they also have a strong need for personal archiving and disaster recovery. The importance of backup and preservation for SMEs and current shortcomings therein are pointed out in [4].

Commercial services for data recovery mostly focus on hardware failure and recovery of data from failed disks, which might not be possible in all cases, or an expensive process. However, the Internet has brought a certain shift in terms of data storage – a few years ago, one could assume users have most of their data stored on hard disks and other storage media, but this assumption does not hold any more. People increasingly use free, virtually unlimited e-mail accounts to store their documents as attachments in messages remotely [1]. Specific types of files might be stored in social media service sites such as Flickr, and in collaborative work sites such as Wikis. An approach to use newsgroup services and mail attachments as back-up strategy is described in [3]. However, it requires that the data is deliberately spread across external sites, rather than relying on the inherent replication in personal mail or groupware services.

In this work, we rather focus on the scenario that a user is suddenly left without his or her main file storage on his desktop computer and no backups exist or that they only comprise a very limited set of data. We emphasise methods

M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, pp. 455–458, 2009.

for analysing, mining and recovering from online, distributed data storage that do not require planning beforehand. Structuring huge amounts of attachments so that they resemble a usable structure is a challenge, requiring sophisticated analysis of the objects and especially their relations to each other. Many sources provide contextual information in terms of people involved in the communication, e-mail text, blog entries, or various comments. We thus aim at identifying the context of objects used in the same *time*, within the same *project*, with similar *content*, with a certain set a of *people*, and of a certain *type*. We present a prototype that offers visualisations of such context, and allows users to interactively organise their objects into meaningful folders.

2 Establishing Context of Digital Objects

Our application scenario is disaster recovery, i.e. to recover as many files as possible after a data loss. Rather than aiming at exact reconstruction, we organise the objects in a new, semantically meaningful way, allowing the user to adjust the automatically established structure. We focus on the following steps.

Object Recovery. We extract all digital objects from the various data sources, such as cooperative work platforms as BSCW and Wikis, e-mail accounts and other sources. This step is rather simple, and requires tools using the correct protocols to access remote information, and processing rules to identify objects worth harvesting.

Context Extraction. We then automatically extract semantic contextual information of objects. Context is present in several forms, ranging from low-level technical context in which the object was created, via its immediate context of use, such as people involved or the activity it is related to, up to a wider sociological or legal context. In this work, we consider only context that can be established in a semi-automatic manner, along several partially orthogonal dimensions. The principle of using various different dimensions is inspired by the concept of data warehouses and on-line analytical processing (OLAP). While the number of potential dimensions that digital objects can be organised by may be larger, we currently use the following in our first prototype: (1) the *time* of object creation, modification and use, (2) the content/file *type*, (3) the *people* involved, and (4) the *content*, across different sub-categories, such as *the topic*, *the genre*, *acronyms*, for example in project names (consult [2] for more details).

Combining Context Dimensions and Storage. These dimensions can well be used separately, but the true potential lies in combining and contrasting them with each other. Combining the temporal and content/project dimensions can for example identify sets of documents of various file formats that belong to one 'instance' of a yearly-repeating project reporting – slide-shows from the presentation, spreadsheets detailing financial aspects, and text-documents describing the outcomes. Using the time-dimension only, we could not figure out which documents belong to the same project, while the file type dimension would wrongly

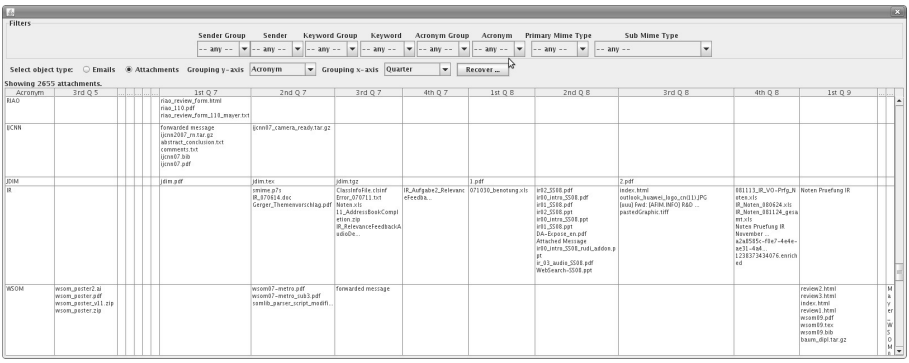

Fig. 1. Pivot-table View on E-mail Attachments

separate the documents. Analysing data in a data warehouse is often performed using pivot-tables, which allow to e.g. see all the sales for a specific city for a certain product at various different levels of aggregation. We provide a similar tool to combine and contrast the relation of objects on those orthogonal dimensions, depicted in Figure 1. It allows users to organise the digital objects along certain dimensions deemed important, by grouping the objects along two different dimensions, and filter the documents along the same or other dimensions.

3 Experiments

We chose e-mails for initial evaluation, due to them being widespread, easily accessible, and safe from local data loss because stored on remote servers. Even though some e-mail corpora are available, we could not use any of those for the lack of attachments in most of the publicly available ones. We thus used two personal mailboxes of the authors of about 19.000 and 23.500 e-mails, from early 2005 and 2006 until early 2009, resp. 2.310 and 1.287 e-mails contained a total of 2.515 and 5.923 attachments. One important aspect in e-mail communication is that often a set of documents get sent packed in an archive format, e.g. as 'zip' files, while in a user's home directory, they are often in their original form. Thus, we also consider the files contained in archive files. The mailboxes exhibit similar structures: both were used primarily for work-related communication, mostly with people inside the same group, European and national project partners, students, and few private e-mails. Topics cover e.g. scientific publications and reviews, project management, and communication with regard to teaching. Differences between the mailboxes are mainly in the focus on different projects and conferences, but not in result performance, thus we focus on the first user.

An exemplary disaster recovery process related to scientific papers could be done as follows: (1) select the time and acronym dimensions for the grouping, (2) select only those acronyms that denote conferences, (3) combine cells on the date dimension to correctly identify the various time periods for each paper

submission to each conference in each year, and optionally (4) mark only those cells that shall be exported, for example, de-select the cells 'ECDL / 2005' and 'ISMIR 2007', which rather denote materials for organising the conferences, and finally, (5) start the export by creating 'papers' as the parent folder; the system will assign each cell to a separate sub-folder, which is named by the acronym and date. After each step, we can choose to have the files just exported removed from, or kept in the list of available files. Thus, the set of documents to be stored can be reduced step by step, also facilitating the analysis process. Alternatively, objects can be recovered redundantly in several locations.

For a quantitative evaluation, the number of files recovered from the mailboxes were compared with the respective home directories of the users. To facilitate initial analysis, we opted for a sub-set of the most *relevant* folders, selecting those to data files, and omitting those that have separate back-up regimes, such as programming/development parts that are covered via version management repositories. The total file-count in this directories was 10,000, of which we further skipped temporary files. This resulted in 5,500 files, out of which we could match 1,730 documents (31%) with the files on the specified home directory folders. This folder still contained a lot of 'bulk' files, e.g. some data used for experiments. Therefore, we selected folders containing all the files of the user related to conference papers, peer reviews for conferences, teaching, and project proposals and reporting. After applying the same filtering, the total file-count in these directories was 3,000, of which we could recover 1,250 files (42.5%).

4 Conclusions and Future Work

We presented analysis tools to assist users in recovering their files after a data loss. We recover files from online repositories, and then semi-automatically establish the context of and relationship between those digital objects and allow for analysis based on pivot-tables. The experiments in this work showed promising results in the academic domain. Future work will thus focus on testing our approach in other settings, like small and medium sized enterprises, as well as evaluating the approach focusing on user satisfaction.

References

1. Marshall, C.C., Bly, S., Brun-Cottan, F.: The long term fate of our digital belongings: Toward a service model for personal archives. In: Proc. IS&T Archiving 2006 (2006)
2. Mayer, R., Rauber, A.: Establishing context of digital objects' creation, content and usage. In: Proc. Int. Workshop on Innovation in Digital Preservation (2009)
3. Smith, J.A., Klein, M., Nelson, M.L.: Repository replication using NNTP and SMTP. In: Gonzalo, J., Thanos, C., Verdejo, M.F., Carrasco, R.C. (eds.) ECDL 2006. LNCS, vol. 4172, pp. 51–62. Springer, Heidelberg (2006)
4. Strodl, S., Motlik, F., Stadler, K., Rauber, A.: Personal & sme archiving. In: Proc. 8th ACM IEEE Joint Conf. on Digital Libraries (2008)