Interacting with (Semi-) Automatically Extracted Context of Digital Objects

Rudolf Mayer
Institute of Software Technology & Interactive Systems
Vienna University of Technology
mayer@ifs.tuwien.ac.at

Robert Neumayer
Department of Computer and Information Science,
Norwegian University of Science and Technology,
neumayer@idi.ntnu.no

Andreas Rauber
Institute of Software Technology & Interactive Systems
Vienna University of Technology
rauber@ifs.tuwien.ac.at

ABSTRACT

The context in which digital objects are created, modified, or used is essential for the interpretation of information entities, for retrieval settings, for establishing their authenticity, as well as ensuring appropriate use. Therefore, determining this context of creation and use of digital objects is an essential task for many areas and applications, from (huge) digital library settings to end-user applications such as search. However, context is notoriously difficult and laboursome to establish and document, and when it has to be entered and maintained manually by the creator of the digital objects, it is often missing or partially incomplete or incorrect. Thus, this paper proposes an approach to (semi-) automatically determine the context of creation and usage of digital objects. Various facets of context along different dimensions are automatically detected, and are combined in pivot-table inspired views, at multiple levels of granularity, which then allow the extraction of the most appropriate connections to other digital objects. Finally, this contact can be used for a range of applications, such as search and navigation.

1. INTRODUCTION

Digital objects, as all kinds of information, do not exist as isolated snippets. Rather, they are embedded in, and also form themselves a larger context, an information space, where context is defined as how digital objects are relating to each other. Various facets of such a context of digital objects exist. It may be the setting and intention within which they were created, the persons and activities involved, or the time frame and other possibly correlated activities that are in some way linked to a digital object, or influence it. Thus, context consists both of the relationship in terms of metadata that a piece of information shares with other information items (such as time period of creation, object type, purpose, creators or users/recipients, and others), but also of the embedding of the the very content itself, that a piece of information is conveying. Here, a piece of information may be both a digital object as such, but also a subset of it, e.g. a certain information item in it, or a content snippet. It may, however, also be applied to a larger group of objects which are already bound together by a common context, such as e.g. an email body and its attachments. Thus, context of digital information basically describes all relationships and commonalities of a piece of information with other information items and dimensions along which these can be structured. A taxonomic description of context for information mediation in Digital Libraries is given in [11], which organises context in *Information context*, *Community context* and *User context*. While we capture some aspects of *Community context*, our primary focus is on *Information context*, which "includes information about the digital information object that is directly related to the individual artifact, and to its surrounding information structures like an information collection it is part of".

All those contextual dependencies are essential enablers for the proper interpretation of digital objects. Moreover, they also provide important clues for identifying relevant pieces of information for a given information need, and as such form a basis for retrieval tasks.

Yet, establishing and documenting the context of information objects is a notably difficult and time-consuming task. Organisations collecting a lot of objects, such as cultural heritage institutions or archives in industry settings, press agencies, etc. have a primary interest in having those objects well described. Thus, they take great care and invest substantial effort in correctly and extensively documenting the various types of context of individual pieces of information. For the need of having an as extensive documentation of the context of digital objects as possible, and given the tremendous effort required to establish and document it, a range of approaches has been developed that aim at automatically capturing context. This may range from simply documenting essential metadata as part of the creation process of a document, for example author and time information as found in standard office applications, to more complex workflow environments, where groups of related objects are cross-linked, bundled, and the various stages in their creation and usage are documented automatically.

In less integrated or controlled environments, however, documenting the context of information objects is in most cases outright neglected. This concerns many small and medium enterprises, as well as small home and office envi-

Proceedings of the 1st Workshop on Context, Information and Ontologies, June 1, 2009, Heraklion, Greece

ronments. On the one hand, this poses problems for the very people creating and using the content, who are finding it increasingly difficult to locate certain pieces of information, i.e. a specific photograph, a letter, or a certain e-mail discussion. On the other hand, it constitutes a significant challenge in professional settings such as archives and libraries, when receiving donations or bequests e.g. from famous artists, to be ingested in a structured way into their collections.

Completely and correctly establishing all facets of the context of a digital object may never be a fully automated process. Yet, a range of techniques may be applied to reduce the level of complexity. For example, it can be of great help to identify potential correlations between objects, events, persons, or activities across time to semi-automatically establish a basic context of pieces of information in a bottom-up manner. Context between digital objects can be present and established on several different and orthogonal dimensions, for example objects created in the same *time*, within the same *project*, with similar *content*, with a certain set a of *people*, and of a certain *type*.

In this paper, we thus propose methods to (semi-) automatically detect and recover inter-object context in multiple dimensions. This automatically established can then be refined and adjusted by the user in an interactive process. We further describe a research prototype which aims at analysing various different information objects across several dimensions, with the aim to identify certain patterns of relation between them. At the moment, this prototype is focusing on e-mail conversations and attachment contained therein, but support for other repositories is under way. The prototype relies on techniques from machine learning, information retrieval (specifically natural language processing) and uses concepts found in online analytical processing, to detect groupings and correlations between objects and offering respective views on an otherwise indiscriminative object collection. By combining multiple views, certain objects can be related to each other in a semi-automatic manner.

This thus established explicit context information may be utilised as essential finding aids in personal information repositories, but may also be used in professional settings, for example to enhance metadata descriptions beyond those commonly used. The contextual information can also be employed in a range of other applications. One example could be disaster recovery, when after the loss of a user's home directory, missing elements are to be recovered from other external repositories, such as e-mail inboxes or cooperative work platforms, and structured in a semantically meaningful way.

The remainder of this paper is structured as follows. Section 2 gives an overview on related work. Section 3 then introduces the different types of context we consider for our (semi-) automatic extraction and detection process, while Section 4 outlines some possible application scenarios for utilising the newly generated context. Section 5 then presents experimental results on two personal repositories, and shows the feasibility of our approach. In Section 6 we give a conclusion and outline directions for future work.

2. RELATED WORK

The management (and preservation) of digital objects has become a major research objective in the field of Digital Libraries. Along with metadata, which provides additional information, the relations between the objects and their context, such as the setting in which they were created, are additional facets to consider. For a news article, for example, knowledge about its setting, time, and persons involved in its creation, means of distribution, and discussion on it in other media, may be crucial for a better understanding and fair assessment of its authenticity and reliability. Articles published on private blogsWeb pages have a whole different impact than official news bulletins. Such information is seemingly easily assessed by humans, the automatic integration into end-user applications for searching and organising their documents or for digital library systems, however has proven to be difficult and identified as a major task for the future [12].

The importance of context for digital objects is discussed in [15], focusing in particular on the specificities of *digital literature*. It is pointed out that, in addition to the technical environment a digital object was created in, also the sociological environment should be preserved. This is especially important for volatile digital objects, e.g. in networked or pending writing, which involve several authors working collaboratively, both being actively involved during the creation step. Similar issues can be found in collaborative tools such as a Wikis.

Social networks and their user data have been studied in a variety of contexts ranging from specifically designed online communities to automatically constructed networks from e-mail data. For example, various publications have been based on the Enron corpus[1]. This corpus was also studied in [5] showing the applicability of social network analysis to real-world scenarios as well as problems therewith. However, this corpus is not feasible for our analysis since attachments were not provided.

A number of tools and services have been developed that perform content characterisation. The National Library of New Zealand Metadata Extraction Tool[2] extracts preservation metadata for various input file formats. Harvard University Library's tool JHove[3] enables the identification and characterisation of digital objects. Collection profiling services build upon characterisation tools, such as DROID, and registries, e.g. PRONOM [3], to create profiles of repository collections.

In addition to the predominantly metadata-oriented information extracted by the tools introduced above, text mining techniques are employed to extract content-specific information. Natural language processing tools such as GATE [4] allow the detection of concepts such as names, places, dates, or part-of-speech structures such as pronouns, adjectives, etc. Gate relies on a range of thesauri as well as rules to identify these concepts.

Once a myriad of characteristics has been extracted from digital objects, these need to be analysed in order to identify patterns, first individually within each of them, then by combining them to identify correlations. To this end, machine learning techniques are employed to analyse either the cluster structure or to assign category labels to objects. Specifically, Wards clustering [7] as a representative of the agglomerative clustering algorithms is used to identify patterns, and provide hierarchical organisation. A detailed discussion of machine learning and its use in text categorisa-

[1] `http://www.cs.cmu.edu/~enron/`
[2] `http://meta-extractor.sourceforge.net/`
[3] `http://hul.harvard.edu/jhove`

tion, a subtask of text mining, is given in [16].

While most content analysis in this domain focuses on the detection of topical aspects, the functional characteristics of documents tend to be at least as important in assisting in their correct interpretation. This can be addressed by performing genre analysis of the digital objects, extracting textual, structural and layout features [2] as well as specific keywords, that can be used to train a classifier sorting documents into specific genre categories such as minutes, memos, papers, homepages, notes, listings, etc. [9, 8].

When recovering objects from e-mail archives, means for structuring the resulting object collection and navigating it are essential. A range of different visualisations in the context of desktop search are evaluated in [6]. Its user evaluation showed that Tree View and Cloud View are the most useful visualisations for presenting search results. The Tree View takes into account user defined folder structures pointing out their feasibility for document organisation purposes, which further motivates the research presented in this paper.

The approach of analysing digital objects introduced in this paper is based on concepts present in data-warehouses[10]. These data repositories store vast amounts of typically numerical business transaction data, as well as dimensional data, which are reference information used to bring context to the business data. Analysing such data is often performed using the online analytical processing (OLAP) approach, allowing for fast multi-dimensional queries. A central concept is the OLAP cube, which prepares the data for fast analysis. Data warehouse and OLAP principles have been further applied to less-structured objects, forming so-called information warehouses. For example, such systems may focus on the analysis of web-pages, as e.g. presented in [1] and [13].

3. (SEMI-) AUTOMATIC CONTEXT GENERATION

In our approach, we consider several different types of digital objects, such as files in the user's directories and external media such as CDs or DVDs, blog entries, Wiki pages, or e-mail messages.

Context is present in several forms – it ranges from a very low-level technical context in which the object was created, via its immediate context of use (such as people involved, the project or activity it is related to, etc.) up to a wider sociological, legal or cultural context. While all levels of context are of importance for the authentic interpretation and usage of a digital object, we focus mainly on the narrower focus of context to be determined (semi-) automatically.

Therefore, we regard the detection and documentation of context of digital objects as a semi-automatic process along several different and partially orthogonal dimensions, each of which structures objects according to different aspects. While the number of potential dimensions that digital objects can be organised by may be larger, we currently use the following dimensions in our first prototype:

- the *time* of object creation and modification
- the content/file *type*
- the *people* involved
- the *content*, across different sub-categories, such as
  - the topic
  - the genre
  - acronyms, for example in project names

Other dimensions, not yet included in the current prototype, comprehend references to places and other more detailed semantic concepts, existing structures in the document repository, e.g. directory structures, e-mail folder structures and e-mail conversation threads, as well as specific object characteristics, such as embedded metadata (for example EXIF headers for digital photographs), for which specific extractors will need to be integrated into the system.

These dimensions can be used separately and independently of each other, or be combined. This is useful, for example, to establish the temporal context of objects generated periodically, for instance, yearly repeating process of reporting for the previous year. Analysing the objects using only the project dimension, it is not immediately obvious that objects from two different periods might not be that highly correlated. Grouping information objects by time alone allows for easy discovery of for example the actual focus of work in a specific time. Yet, those objects might stem from several independent activities and might thus not otherwise be related. Thus, different dimensions need to be combined to establish a more precise context.

The principle of using various different dimensions as orthogonal views on the data is inspired by the concept of data warehouses and online analytical processing (OLAP) [10]. Data warehouses are (huge) databases storing vast amounts of general (numerical business transaction) data, and dimensional data. Typical examples for dimensions are time, geographical information, customer information, etc. Often, these dimensions can be organised in hierarchies, to provide contextual information at different levels of detail and aggregation. The time dimension for example could be organised as certain periods of a day, the weekdays, weeks, months and years; geographical dimensions might be aggregated from boroughs to cities, states and countries. Organising a data warehouse with a central fact table and hierarchical dimensional tables is called a 'Star' or 'Snowflake schema'. A graphical representation of the latter is given in Figure 3, depicting some of the dimensions we present in this paper.

Analysing data in a data-warehouse is often performed using the online analytical processing (OLAP) approach, which allows for fast multi-dimensional queries. A central concept is the OLAP cube, which prepares the data for fast analysis. The analyst can pivot the data in various ways, e.g. see all the sales for a specific city for a certain product, and do this at various different levels of aggregation, allowing easily to get more detail on demand ('drill down') or a more abstract, summarised view ('roll up' or 'drill up').

Applying these concepts to the process of establishing context for digital objects, we want to describe the dimensions identified in more detail. It has to be noted that the degree of context establishment automation varies among the several dimensions, from fully automated along pre-defined categories, such as MIME file types and the time dimension, to semi-automated e.g. for projects, where an initial hierarchy of projects can be proposed, but which benefit from human interaction to improve the quality. Examples and details of the way these dimensions are used will be provided, alongside presentation of experimental results, in Section 5. Further, the degree to which certain contextual dimensions

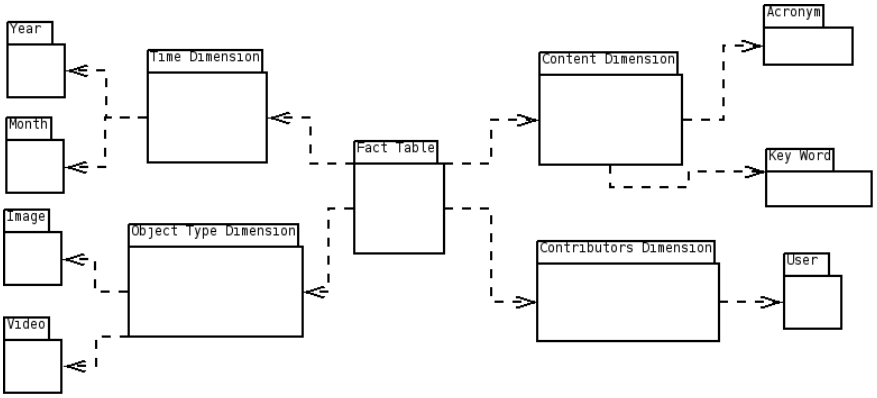

Figure 1: 'Snowflake' schema of a data warehouse for context information. The central table states the fact about the digital objects (e.g. metadata). The dimension tables such as content or type signify the various categories to organise the objects by

of objects can be filled depends to a certain extent also on the source of the digital objects – e-mails for example hold a rich set of clues regarding the people involved in the use of a specific digital object, while objects found in a home-directory indicate the user's personal preferences of organising them.

3.1 Time Dimension

The time-dimension can be defined very much in analogy to the time dimension in conventional data warehouses. For example, it could be structured as follows: 'hour of the day', 'day of the week', 'week', 'month', 'quarter', and 'year', with 'week' forming a separate aggregation branch in the hierarchy, crossing the boundaries of months, quarters and years. This dimension can be very easily and quickly customised to the precise use needs.

3.2 Object Type Dimension

Digital objects can further be distinguished by their file type. One categorisation for file types is the 'Internet media type', sometimes also called 'MIME-type' or 'Content-type'. This categorisation is a de-facto standard for object exchange over the Internet; it is suited for our task, as it provides a, even though limited, hierarchical order on digital objects, in the so-called (primary) 'type' and 'subtype'. Primary types are among others 'text', 'audio', image', 'video' and 'application', while subtypes can detail the content further, e.g. as 'application/pdf' or 'text/rtf'.

Several tools exist to correctly identify the MIME type from an object, along with a wealth of characterisation tools that provide more detailed information, such as the aforementioned DROID Digital Record Object Identification tool in combination with registries such as PRONOM [3].

3.3 Contributors Dimension

The people involved in creating or editing digital objects may indicate a relation between those objects, especially if the same group of people are working together on a set of objects that are otherwise not related to each other e.g. in the type dimensions. These groups might then indicate organisational units such as companies or departments, or project teams across different organisations.

To populate this dimension, we need to first identify the persons involved in the object creation process. For several file formats, this information is stored as explicit metadata in the object itself; for other objects, this information may be derived from its storage repositories – Wiki pages or revision control systems such as CVS or Subversion, for example, provide a complete change history. Another example is e-mail, where sender and receivers can be put in context with digital objects attached to other messages – often the creator of a document is the initial sender of an attachment.

We then perform identity detection, to correctly map senders and recipients to persons. We can further establish links across different digital object repositories, for example by resolving the name specified in the 'author' metadata field of a Word-document to the same person that we have received e-mails from and that edited a Wiki page. In the latter case, the person may not be identified with his or her name, but a specific user-name, but identification can still be possible via the e-mail address associated with the user name. Some simple heuristic resolution steps can be done automatically, while others can be done semi-automatically, with the system proposing likely matches and the user interactively confirming them. Instead of a flat dimension we want to introduce some hierarchy, e.g. in the form of people working together at different levels. We thus perform clustering of the persons involved in the object creation process, simply regarding the persons as our features which describe our object instances. Ideally fitting for this task are algorithms that produce a hierarchy of clusters of persons, thus allowing to view their relation at different levels of detail.

3.4 Content Dimension

Information objects that share some similarities in their content with a specific other object form the context of that object. Content relation can be detected on several different aspects – the usage of similar keywords, similar style, etc.

The topical or content dimension of digital objects primarily refers to the plain text of a document itself, or text documents which describe the object more closely. In the case of text documents the content can directly be indexed; for non-textual attachments, e.g. the e-mail body or Wiki page containing the attachment may serve as a surrogate. This method has the advantage of 'creating' context for all types of attachments – whether they could be textually indexed themselves or not, as e.g. image or video data.

Different representations of textual content can be considered for analysis, e.g. standard bag-of-words based full term indexing [14]. Apart from this, more advanced indexing techniques are utilised to identify specific patterns and types of content. Other aspects that can be extracted using natural language processing tools such as GATE, are names of persons, places, or dates. Such indexing commonly results in a vectorial representation of documents that can be analysed using machine learning techniques.

Further, we stress the importance of 'within-project' similarity as a special case of content similarity and vital aspect of an objects content – digital objects created for and within the same project share a strong contextual relation. Projects might be characterised by dealing with a certain topic; however, that assumption might not hold for larger projects with several independent tasks. Further, it is difficult to label such topical relationship with meaningful, unambiguous terms. Therefore, another approach to identify projects is simply to detect project names; very frequently, these names are in the form of acronyms, thus project identification may be reduced to acronym detection. However, acronym detection is a difficult task on its own, as in many cases, acronyms are existing words, which makes detection more difficult. Once acronyms are detected, we can, similarly to the dimension of contributors, automatically cluster them to achieve a hierarchical ordering of their relations. Also, acronyms can be used to further structure existing clusterings based on text content.

4. USING OBJECT CONTEXT: APPLICATION SCENARIOS

As aforementioned, the context of digital information objects is essential to to correctly interpret and use information. Thus, such automatically established context of digital information objects can be utilised in a multitude of tasks, especially in settings where the context oj such objects is not obvious to the user. This specifically affects large archival holdings, where multiple users are creating, modifying and using digital objects. It also affects enterprises and home offices, as well as private users, as they all accumulate significant amounts of information. This can extend to a degree where they lose oversight, and are unable to find relevant pieces of information for certain tasks. Or, they might forget which of the several different versions of an object is the current one, and whom it was sent to. Thus, the following application scenarios could be improved by enriched object context.

Search and navigation.

In an everyday tasks, contextual data can help in finding, opening, storing, or accessing similar objects. One simple possibility is suggesting target folders as a context menu when saving a file from an application. Exploiting context information gathered as in our prototype, the system might be able to provide accurate suggestions for storage locations. For example, an attachment from an e-mail received from another member in a certain research project should probably be saved along with other relevant documents for this project, maybe considering only those folders already containing objects received from that specific person.

Another primary task can be search support. Most search algorithms currently only analyse the textual content of an object, and thus retrieval is limited to objects containing the terms specified in a query. This is of limited use for identifying predominantly non-textual objects such as multimedia data (images, video, audio), or even textual objects that sometimes may not match the query terms. This might hold true especially for digital objects with only limited quantities of text, such as slides for presentations. With contextual information at hand, the initial search results can be expanded by objects sharing a very similar context as the retrieved ones.

Further, cotext-based browsing of objects could provide a complementary view on collections, allowing a different kind of navigation, thus augmenting and enhancing folder-based navigation and keyword-based search.

Moreover, specialised search tools in applications can be enhanced by contextual information. The study on e-mail documents presented in the next section indicates that e-mail clients could greatly benefit from this kind of context-inspired search, rather than using traditional keyword-based search.

Disaster Recovery.

A frequent incidence, especially in less professional computing environments, is the loss of significant amounts of data, caused for example by hardware failures or loss/theft of laptops. Often all information on a users home directory or personal harddisk is lost, with only partially or heavily outdated back-ups available for data recovery. However, increasingly a rather large fraction of objects has been sent or received via external communication channels such as downloads from project websites or, specifically, via e-mail. These objects are thus usually stored on external systems, and thus frequently not affected by the local disaster. Recovering these massive amounts of information manually tends to be an almost unmanageable endeavour. A goal for disaster recovery assistance would there be to automatically extract and group these objects by means of contextual information, thus restoring a part of the objects on a user's harddisk, and structure them in a logical waythat can feasibly be used.

Object ingest for Digital Preservation.

A different scenario in a more professional setting relates to ingesting large quantities of rather unstructured information in archival institutions for long-term digital preservation. Identifying which objects are contained on a number of data carriers in a box and how they relate to each other, what they are dealing with, and when they were created, constitutes a rather tedious task that can be supported by advanced tools that help to identify and suggest potential contexts often occurring in digital memory institutions. This frequently occurs in archival settings when institutions accept donations of digital materials. At the same time, such a solution may assist small institutions and home users to move their ad-hoc 'curation' of their digital objects to an improved level by helping users to collect and maintain context information as part of a small office or home archive [17].

5. VIEWS OF CONTEXT FOR ANALYSIS OF E-MAIL CORPORA

To demonstrate the methods and approaches described earlier, we carried out several experiments on different subsets of digital object collections, with a focus on real-world

e-mail archives. These are combined with the users (home) directories, and are expandable by other data sources, such as Wikis, blogs, and external media such CDs or DVDs. One typical application scenario is to find back all objects related to a certain event.

Context extraction from e-mail archives as such is a rather new area of research and thus lacks available test corpora. E-mail corpora in general have been used for a long time with a special focus on junk mail detection. However, these corpora are not feasible for the experiments performed in this context and we decided to use new corpora. This is especially motivated by our experiments focusing on a subjective evaluation, which would be very difficult to perform to the same extent on unknown or less-known corpora. Thus, we focus on personal e-mail inboxes, where no additional preprocessing is applied except for the removal of spam folders. We carried out experiments on the inboxes of two of the authors, which contain approximately 18.000 and 23.500 e-mails, resp. Out of those, approximately 5.300 and 6.500 were e-mails sent by the authors. The collection covers a period from the early 2005 and 2006 until early 2009. 2.310 and 1.287 e-mails contained a total of 4.605 and 5.923 attachments. The e-mails were written in different languages: for the first collection, 13.800 e-mails were in English language, the remainder was in German; the second collection contained 13.700 German messages. Both e-mail accounts were used primarily for work-related communication, mostly with people inside the same group, European and national project partners, and students; additionally, both inboxes contain a smaller amount of private e-mails. The e-mail inboxes were combined with the respective home directories of the users.

5.1 E-Mail as Extended Text Documents

As opposed to plain text files, e-mail data offers a wealth of possibly useful contextual information to exploit. Not only are e-mails divided in *from*, *to*, *subject*, or *body* fields, they also offer attachments as special purpose files which can be brought in context with the data found in the rest of the e-mail they were originally sent with or additional information mined from the e-mail corpus in question, as well as in the local file system or objects in other repositories, such as revision control systems or collaborative tools such as a Wiki or BSCW.

5.2 Context Dimensions

As e-mail documents have the above-mentioned specific properties and implicit metadata, some of the dimensions can be developed in more sophisticated ways than it would be possible with digital objects from other repositories. In detail, the dimensions are populated and utilised as follows.

5.2.1 Time Dimension

The dimension is simply generated using the (sending) 'Date' field of the e-mail. The system then allows to view the digital objects at a customisable grouping, by selecting levels of aggregations, e.g. a grouping per month or year. For files in the user's directories, we can analogously use creation and modification dates.

5.2.2 Object Type Dimension

The e-mail 'body', i.e the textual content itself, is always of the MIME type 'text', generally of the subtype 'plain', whereas sometimes a rich-formatting such as 'html' is used; thus, there is little value in providing specific views on the e-mail body. More interestingly, though, the attached files can be of any file type. We thus detect their MIME types, and can arrange them in a browsable tree. In the case of using the MIME type, this hierarchy consists of two levels, but other object characterisation tools that offer a classification with a different hierarchy might be easily employed instead in our system. With this view, it is thus easy to find files of a specific type, e.g. it is easy to find image files; additionally, with the hierarchy, we will not only find files of one specific subtype, but all objects of the same type will be closely located to each other. The same principle is applicable to the files in the home directory.

5.2.3 Contributors Dimension

For the subsequent processing steps, we can enhance the quality of the context by first resolving duplicates from the collection of digital objects. A duplicate in our context is a user contributing to the authorship of digital objects under more than one personal name (e.g. the name info set in e-mail clients); this can e.g. be just a different spelling of the name, such as mixing up the order of first name and surname, or more complex issues such as having a different username for the Wiki and the Source Code Repository. Obviously such duplicates can severely impact the outcome of any data mining and context detection tasks performed. If the personal address on a person's home computer is set up slightly different from the one at the office workstation, a certain amount of contributions will be under this second name. As a first step we identify persons using multiple personal names for the same e-mail addresses. We then substitute each personal name with the one most often used. This alone filters out many authors with duplicate addresses and finds an author's main address or personality. Then, we identify persons operating with multiple addresses and unify these based on substitution rules. Although rather simple, these rules reduce the amount of duplicates to a satisfactory level and exhibits relations in a better way.

E-mails exhibit rich information on people involved in the same conversations in their sender and recipients fields. We can thus utilise this information to construct a social network graph of people related to each other, to allow the user to analyse the relationships of contributors in his repository of digital objects. In the specific case of an e-mail repository, the connections in the graph are defined by people being among the senders or recipients of the same e-mail exchanged, while the strength of the connection is the count of the messages exchanged. Similarly, such a graph could be constructed for contributors of a Wiki, or any other digital object that stores authoring information, e.g. by means of co-authorship.

The graph can then be displayed, explored and searched interactively, as depicted in Figure 2. As additional analysis aid, we provide highlighting mouseover effects which mark the selected person and highlights all the connection, a helpful tool for detecting sub-graphs of people closely related. Further, the user can control the level of detail with a variable threshold for the connection strength. This view thus provides a good starting point for analysing social structures among the authors of the digital objects. Further, the visualisation allows for interactively editing the identified persons, e.g. it allows to 'merge' two names that have

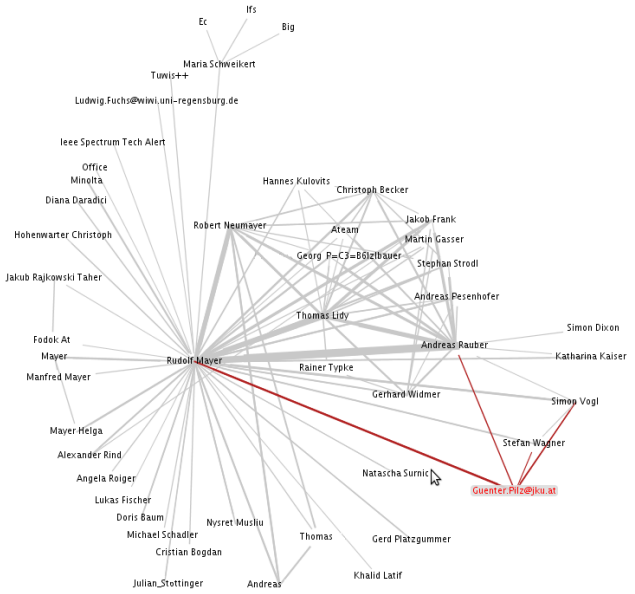

Figure 2: Graph view of the social network

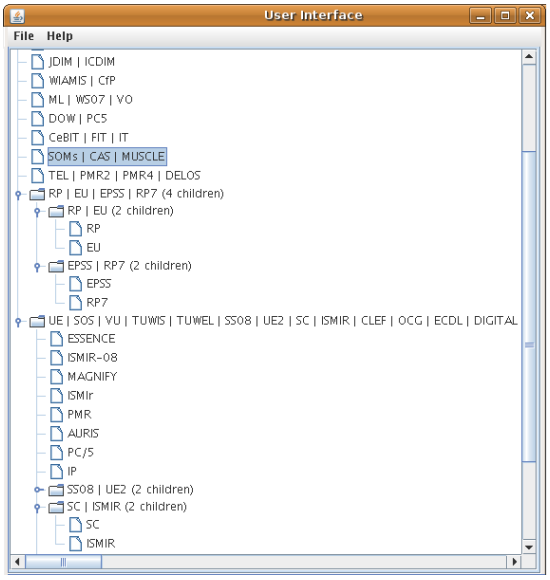

Figure 3: Hierarchical view of projects, identified by their acronyms

not correctly been resolved to the same person, and persists that modification by adding a manual rule to the identity detection.

To provide a hierarchical view of the collaboration, we consider the sender(s) and recipients of e-mail as the features describing it. In the beginning of the subsequently applied agglomerative clustering process, every digital object forms its own cluster; in each subsequent step, the two nearest clusters are merged, until finally only one cluster remains. We employ Ward's linkage as one of the most performant within the linkage clustering families. In this algorithm, the distance of each pair of clusters is defined by the increase in the 'error sum of squares' if the two clusters are to be combined. For more details on the agglomeration computation, please refer to [7].

The result of the Ward's algorithm is a hierarchy of clusters which the user can browse through. Increasing the number of displayed clusters means splitting existing clusters into two new ones, while reducing the number of clusters is achieved by merging two clusters into one. We can further utilise the clustering tree to generate a hierarchical dimension, if we define a number of n levels, with m_n different clusters on that level. From an e-mail repository, we create a matrix of all the persons involved in the e-mail communication in the same manner as for the graph visualisation. We then apply Ward's clustering to detect groups of people collaborating on the creation and modification of sub-sets of the objects.

5.3 Content Dimension

As described in Section3.4, one approach to detect projects is to detect the occurrence of project names, which are often in the form of acronyms. Acronyms are often characterised by being written in uppercase, or at least mostly uppercase, letters. Acronyms are often designed to be pronounceable; while many of these are not dictionary words, there is still quite a number that are actually dictionary words, such as e.g. the 'MUSCLE' or 'PLANETS' EU projects. Consequently, more sophisticated approaches to detect acronyms in a text document, e.g. comparing occurrence frequencies of this term with the terms in a reference corpus, have proven to exclude these kind of project acronyms. Thus, a computationally simpler approach was taken, namely detecting all uppercase and mostly-uppercase terms in the subject lines of e-mails. Subsequently, we applied a frequency-thresholding on the terms, i.e. terms that occurred too seldom were excluded as being not relevant enough, or potentially wrongly capitalised words. The user can interactively adjust the thresholds for the cut-off, to finetune the results for his collection. Additionally, from the list of acronyms proposed by the system as result, he could manually mark terms as non-acronyms, and add other acronyms that were not detected; this process is aided by displaying the e-mail subjects and sender next to a term. With the knowledge of the user on the projects he is involved in, he can quickly scan for wrong terms, and thus have a clean list of correct acronyms within a few minutes.

Once acronyms are detected, both the subject and body of all e-mail documents are scanned for occurrences of the terms, this time not regarding any capitalisation; this is motivated by the observation that often acronyms are not properly capitalised, especially in the e-mail body. Then, we construct a feature vector representation of each acronym, where the e-mails it occurs in are the features. Applying the clustering as described in Section 5.2.3, we can then create a hierarchical dimension of related projects. One example of an acronym clustering result can be seen in Figure 3. We can see a lot of fitting acronym pairs, for example the highlighted grouping of the EU project 'MUSCLE' with a sub-task labelled 'CAS', and one of the technology applied therein, 'SOMs' (Self-Organising Maps). Also the other matches found by the algorithm group related acronym, for example a conference and its ensuing special issue of a journal ('ICDIM' and 'JDIM'), or the course 'Machine Learning' (ML), taking place in the winter semester (WS), and being of the specific course-type lecture (VO).

As an extended use case, we show the identification of digital objects found in both a user's home directory and in his personal inbox. First a list of attachments is harvested from the inbox. Then, we match these results against a crawl of the user's home directory. The intersection of both can either be used to identify most interesting attachments because they are currently in use, or the identification of files present as attachments that could or should possibly be stored on disk additionally. Further, we can exploit additionally structural information, such as the grouping of files in the same file-system or email folder.

5.4 Combining the Dimensions

While the previous sections demonstrated the usability of views on singular dimensions, and their ability to detect the digital objects' context, users can profit from a whole wealth of new object exploration and analysis possibilities on their files when the dimensions are combined together. In data warehouses, the data in the OLAP cubes is often visualised with the help of a pivot table, summarisation tools that can automatically sort and aggregate data from a table, and display the thus condensed information in a smaller, second table. The user can apply filters, which is roughly equivalent to the 'drilling down' concept in data warehouses. Thus, users can fast and in an interactive process change the abstraction level of the data displayed, which can help them to discover different contextual relations between the objects yet unknown to them, and facilitate their search and retrieval tasks.

For an example, Figure 4 depicts a screenshot of the prototype-type pivot-table view. In one of the authors collection, one can find a wealth of e-mails that got assigned to the acronym 'ECDL', the European Digital Libraries conference series. By first filtering those emails by selecting the acronym in the filter box, and then selecting the time as a grouping dimension, the user can very fast realise that the digital objects stem from different 'instances' of this conference, namely once from organising the ECDL 2005, and subsequently for submitting papers to conferences in the following years. Also, the object type dimension can help in this case, as the types of files attached to the acronyms differ greatly, being mainly text documents (LaTeX and PDF) for the authoring activity, and spreadsheets and other office documents for the organisational work.

Another scenario of how this prototype can help in retrieving specific digital objects by using their contextual information would be filtering all objects related with an European Union funded research project with several dozens of partners. Then, selecting the senders/contributors dimension to group the documents along one axis, objects get correctly separated by the different groups of people working on the specific work-packages in the project. Finally, we could select the time dimension as another grouping on the other axis, and would get an aggregated display of information that would reflect the different activities at certain moments in time within certain project subgroups.

Besides allowing interactive search and exploration of the digital objects as presented above, this context could be batch-generated for each document, and the relations to other objects along the several dimensions can be formally and explicitly stored as context metadata of an archival information package in a long-term digital archive. It may also be used to automatically create a directory structure

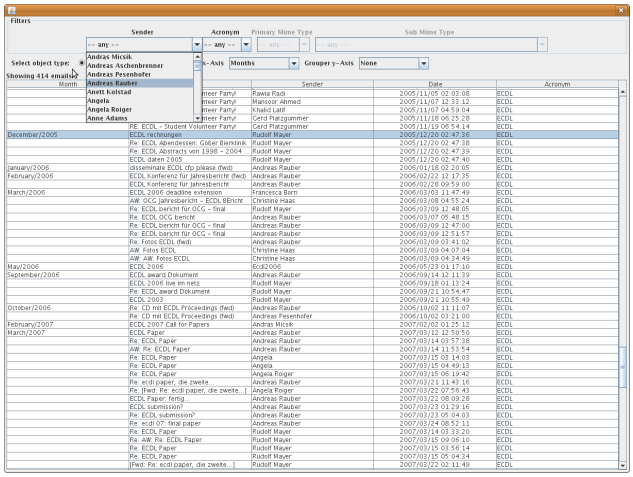

Figure 4: Pivot-table view of emails in the users inbox

for recovery of data from an e-mail archive.

It is important that even though e-mails might exhibit some specific, indirect metadata, such as the senders/recipients, our approach is not in any way limited to e-mails, as similar metadata can be extracted for a wealth of different types of objects.

6. CONCLUSION AND FUTURE WORK

For the proper use and interpretation of digital information objects, knowledge about their context is essential. Documenting this context is a notoriously laboursome task, and this challenge becomes increasingly prominent as digital objects and pieces of information become ever more distributed. In many cases, programs or devices used for creating digital content, such as digital cameras, office applications, etc. already capture a set of context information automatically, such as location information where an image was taken, or information aboutt he user creating an office document. Moreover, applications supporting complex workflows document the process and types of usage of digital objects in a specific environment. Yet, in many cases essential aspects of context have to be documented manually, often (long) after the documents were created, for example when they are to be ingested into a central repository.

To alleviate this process, in this paper we thus presented a framework and a prototypical software solution that assists users in establishing the context of digital objects. This is achieved by analysing the relationship of digital objects across a number of different dimensions. To this end, a number of tools and approaches from the information retrieval and machine learning domain are employed to detect correlations between digital objects according to the people using them, by their content in different degrees of abstraction, from acronyms to full-text indexing, or their types. Once these relations are detected, an interface inspired by pivot-tables used in online analytical processing in data warehouse facilitates automatic as well as semi-automatic analysis of groupings along to those dimensions. This assist users in identifying which digital objects were of interest to which groups of people at which points in time, which in turn allows to establish a certain context of use via a sequence of

objects. Multiple views at different levels of detail help in establishing different types of context for each object, be it a file in a home directory or on a CD, or even stored remotely on Wikis, an e-mail attachment, or being part of a discussion thread or blog posting.

The contextual information thus gained can augment existing metadata of the digital objects, and can be utilised as a finding aid, supporting to a certain degree semantic search. Objects can be retrieved because they are obviously related to a certain aspect, even when the initial keywords in a query do not match the content of a specific object. As such, photographs related to a project meeting can be identified together with other material from the meeting, by correlations in the temporal and recipient domain if they are distributed via e-mail or a shared web storage folder.

The protoype implementation presented in this paper has revealed the potential of this approach. Yet, additional tools that extract more specific metadata for various object types need to be integrated into the system, to fully exploit its potential. Additional sources such as Wikis and other online collaborative tools are to be integrated, extracting more source-specific metainformation. Finally, improvements are required in terms of the system's usability, allowing more intuitive and flexible interaction when analysing correlations across various dimensions.

While the work presented in this paper focused primarily on establishing context at various levels of detail, future work will will focus on qualifying the type of context and adding interpretation by analysing the semantic concepts between relationships, according to the dimension they occur in, at least by providing suggestions for interpretation to be manually confirmed. For example, these could be versions of a document that otherwise has the same file name but increases in size, or discussions on a certain document if references to it appears in the content of another document.

Acknowledgements

Part of this work was supported by the European Union in the 6th Framework Program, IST, through the PLANETS project, contract 033789.

7. REFERENCES

[1] S. S. Bhowmick, W. K. Ng, and S. Madria. Web schemas in whoweda. In *Proceedings of the 3rd ACM Int. workshop on Data warehousing and OLAP (DOLAP '00)*, pages 17–24, McLean, Virginia, USA, November 10 2000. ACM.

[2] D. Biber. A taxonomy of english texts. *Linguistics*, 27, 1989.

[3] T. Brody, L. Carr, J. M. Hey, A. Brown, and S. Hitchcock. PRONOM-ROAR: Adding format profiles to a repository registry to inform preservation services. *Int. Journal of Digital Curation*, 2(2):3–19, November 2007.

[4] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. 40th Anniversary Meeting Association for Computational Linguistics*, 2002.

[5] J. Diesner, T. L. Frantz, and K. M. Carley. Communication networks from the enron email corpus "it's always about the people. enron is no different". *Computational & Mathematical Organization Theory*, 11(3):201–228, October 2005.

[6] S. Foo and D. Hendry. Desktop search engine visualisation and evaluation. In *Proceedings of the 10th Int. Conference on Asian Digital Libraries (ICADL'07)*, pages 372–382, Hanoi, Vietnam, December 10-13 2007. Springer.

[7] J. H. W. Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, March 1963.

[8] J. Karlgren. Stylistic experiments in information retrieval. In *Natural Language Information Retrieval*. 1999.

[9] Y. Kim and S. Ross. The naming of cats: Automated genre classification. *Int. Journal for Digital Curation*, 2:24, 2007.

[10] R. Kimball and M. Ross. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Second Edition)*. Wiley, 2002.

[11] E. Neuhold, C. NiedereÃl', A. Stewart, I. Frommholz, and B. Mehta. The role of context for information mediation in digital libraries. In Z. Chen, H. Chen, Q. Miao, Y. Fu, E. Fox, and E.-P. Lim, editors, *Proceedings of the 7th International Conference on Asian Digital Libraries (ICADL 2004)*, volume 3334 of *Lecture Notes in Computer Science (LNCS)*, pages 133–143, Heidelberg, 2004. Springer.

[12] G. S. Pedersen, K. F. Christiansen, and M. Razum. The use of digital object repository systems in digital libraries. *D-Lib Magazine*, 14(11/12), Nov/Dec 2008.

[13] A. Rauber, A. Aschenbrenner, and O. Witvoet. Austrian on-line archive processing: Analyzing archives of the world wide web. In *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2002)*, pages 16–31, Rome, Italy, September 16-18 2002. Springer.

[14] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[15] S. Schrimpf. Long-term preservation of electronic literature. In *Proceedings of the 5th Int. Conference on Preservation of Digital Objects (iPRES 2008)*, pages 29–30, London, UK, September 29-30 2008. British Library.

[16] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[17] S. Strodl, F. Motlik, K. Stadler, and A. Rauber. Personal & SOHO archiving. In *Proceedings of the 8th ACM IEEE Joint Jonference on Digital Libraries (JCDL'08)*, pages 115–123, Pittsburgh PA, USA, June 16-20 2008. ACM.