SEVENTH FRAMEWORK PROGRAMME

ForgetIT

www.forgetit-project.eu

ForgetIT

Concise Preservation by Combining Managed Forgetting and Contextualized Remembering

Grant Agreement No. 600826

Deliverable D3.4

Work-package	WP3: Managed Forgetting Methods
Deliverable	D3.4: Strategies and Components for Managed Forgetting - Final Release
Deliverable Leader	Xiaofei Zhu, LUH
Quality Assessor	Jörgen Nilsson, LTU
Dissemination level	PU
Delivery date in Annex I	M36
Actual delivery date	March 21, 2016
Revisions	7
Status	Final
Keywords	Digital Preservation; Dynamic Information Assessment; Time-aware Information Access; Managed Forgetting; Policy

Disclaimer

Revision History

Date	Version	Major changes	Authors
01.12.2015	v1	Table of Content	Xiaofei Zhu
11.01.2016	v1	Finished the text	(All)
15.01.2016	v1	Formatted. First version ready	(All)
30.01.2016	v1	Sent for Internal QA	Xiaofei
04.02.2016	v1	QA feedback sent	Jörgen
16.02.2016	v2	Version 2 based on QA feeback	Tuan
21.03.2016	v2.1	Version 2.1 final revision	Claudia

List of Authors

Partner Acronym	Authors
LUH	Claudia Niederée, Tuan Tran, Andrea Ceroni, Kaweh Djafari Naini, Nam Khanh Tran, Xiaofei Zhu
DFKI	Heiko Maus, Christian Jilek

Table of Contents

Executive summary

In the previous deliverables of this Work package, we have reported our work on managed forgetting, including research on memory buoyancy as well as first work on the preservation value and the policy framework. In this deliverable, we deepen our work on presrvation value, present our work on the policy-based preservation framework, revisit the dimension of preservation value and relate them to appraisal. Furthermore, we discuss methods and components for preservation value assessment in different scenarios, as well as applications which realise these methods and components. Specifically, we have conducted a number of studies on the problem of estimating the preservation value for images, the preservation value for text, as well as the preservation value for social media. Regarding the preservation value for images, we extended the work reported in [Kanhabua et al., 2015] by including orthogonal information into the selection model, performing an extensive evaluation, involving more users and gathering more photo collections, and so on. Regarding the preservation value for text, we investigated different settings. For example, we aim at estimating the preservation values of text related to an entity of interest, and investigate how much it contributes to a summary of an entity's (life) situation profile. Regarding the preservation value in social media, we extended our work of public text by conducting a study that uses entities as a pivot to evaluate the past news event's preservation value. Furthermore, we discuss applications which realise the managed forgetting methods and components, including Memory Buoyancy for decluttering semantic information spaces and p reservation value calculation in the semantic desktop.

1 Introduction

In WP3, we aim at developing reusable building blocks for managed forgetting, and formulating a set of strategies for establishing preservation and forgetting processes. In previous deliverables, we have addressed several important problems of managed forgetting, including: 1) the state-of-the-art and key research questions related to the managed forgetting concept ([Kanhabua et al., 2013]); 2) the computational model for memory buoyancy ([Kanhabua et al., 2014]); and 3) the preliminary research on strategies and components, as well as a report describing their functionality ([Kanhabua et al., 2015] as well as initial work on preservation value and its dimensions.

In this deliverable, we develop strategies and provide methods and components for preservation value assessment. In particular, in previous deliverable [Kanhabua et al., 2015], we have discussed basic policy-based preservation framework. In this deliverable, we will present how to make the policy-based framework more user friendly, and address the problem of uncertainty attached to the policies.

On the foundational level, we link the concept of the preservation value to the concept of appraisal. In the context of digital archives, appraisal is a process of determining what is worth keeping. In most cases, appraisal refers to a manual process of estimating preservation value. In this deliverable, we report our anticipated methodological insights for automating the selection process.

Moreover, we investigate methods for preservation value assessment in three different settings, including preservation value for images, preservation value for text, and preservation for social media. Specifically, we investigate the scenario of preservation value for images and develop the work in several aspects, including studying the role of coverage in photo selection, involving more users and gathering more photos, etc. In the scenarios of preservation for text, we assess the preservation value of text with respect to an entity of interest, and study how it could be useful for the summarizing of the entity's situation profile.In the scenarios of preservation value for social media, we assess preservation values of news and investigate what a user remembers and what she might want to re-check about these past events.

Finally, we present two applications which realise the managed forgetting methods and components: 1) Memory Buoyancy for decluttering semantic information spaces. In this work, managed forgetting aims to automatically computing the memory buoyancy of a document with respect to the user attention, and documents with highest values will be recommended to user. 2) Preservation Value estimation in the Semantic Desktop. As the Semantic Desktop (SD) is powerful for supporting both organizational and personal knowledge management, in this work, we give details about the preservation value calculation which has been implemented in Preservation Pilot II and an extended one for the PIMO.

1.1 Structure of the Deliverable

The detailed organization of the deliverable is outlined below.

- Section 2 discusses the dimensions of preservation values and appraisal.
- Section 3 reports the techniques developed for preservation value for images.
- Section 4 explains the techniques developed for preservation value for text.
- Section 5 describes the techniques developed for preservation value for social media.
- Section 6 presents the realization of managed forgetting in several applications, including memory buoyancy in decluttering semantic information spaces, as well as preservation value calculation in the semantic desktop.
- Section 7 presents the extended work on the policy-based preservation framework.
- Section 8 summarizes and concludes the deliverable.

2 Managed Forgetting and Appraisal

In the project we have introduced the concept of managed forgetting, which helps the use in making preservation decisions and at the same time creates benefits in the active use. For this purpose we have introduced two types of information value, namely, memory buoyancy referring to the current value of an object and preservation value referring to the long-term value of an information object.

In this section we focus on managed forgetting related to the preservation value, i.e., to the use of managed forgetting for preservation decisions. This aspect of managed forgetting is closely related to the concept of appraisal as it is used in archives [Cook, 2005, Schellenberg, 1956]. The National Archives in UK define appraisal as ”the process of distinguishing records of continuing value from those of no further value so that the latter may be eliminated” [1]. In this context primary value and secondary value of information objects are distinguished. Primary value refers to ”the value to the organization that created them for administrative, legal and fiscal purposes” and supports in ”identifying records of ongoing business value”. The secondary value, which refers to ”the additional historical value to the organization and wider society”. In each case, understanding the value of information is considered crucial for making preservation decisions. In archives, appraisal is typically done manually based on guidelines. Different from appraisal approaches, we aim for an automated way of assessing long-term information value based on a variety of criteria.

Since the computation of preservation value is a complex task, which depends on a variety of factors, we do not believe that there single methods for this. For considering preservation value on a more conceptual level, we therefore use a set of Preservation Value Dimensions described below.

2.1 Dimensions of Preservation Value

The selection of content to include into a long-term personal or organizational digital memory, is a multi-faceted information assessment problem. In our approach, we compute information value (so-called *Preservation Value*), which is used for deciding on what to include into the personal digital memory.

Definition 1 ***Preservation value*** *is a value attached to a resource reflecting the benefit expected from the long-term survival of the resource.*

In the area of multifaceted information value assessment, several valuation methods have been proposed by employing a rich variety of criteria. Many approaches take observed usage in the past as the main indication for information value, i.e., probability of future use [Chen, 2005, Mitra et al., 2008]. This type of information value is highly associated to

[1] http://www.nationalarchives.gov.uk/documents/information-management/what-is-appraisal.pdf

short-term interests [White et al., 2010], which is influenced by a variety of factors that can be roughly grouped in the following categories: usage parameters (such as frequency and recency of use, user ratings, recurrent pattern), type and provenance parameters (information object type, source/creator), context parameters (such as relevance of resources as background information, general importance of topic, external constraints), and temporal parameters (age, lifetime specifications). Existing works on time decay models can, for example, be found in the field of processing data streams [Palpanas et al., 2004] and time-aware information retrieval [Peetz and de Rijke, 2013, Tran et al., 2015a].

The computation of preservation value is certainly a challenging task. It encompasses predicting the future value of a resource and is influenced by a variety of partially situation-specific factors. Therefore, it is not expected that there will be one single method, which can compute the preservation value for all possible situations, even if we just restrict to the personal digital memory. For example, other factors influence the decision, if I want to keep a photo or a Facebook post.

However, we have identified eight high-level dimensions that can be used to categorize the evidences used for computing preservation value. They provide a basis for developing a more systematic approach to preservation value assessment. The list of dimensions has been compiled based on content selection work from literature [Walber et al., 2014, Wolters et al., 2014], own studies in content selection for preservation [Ceroni et al., 2015a] and a study work on appraisal in the archival domain [Cook, 2005, Schellenberg, 1956]. An initial list of dimensions had already been presented in deliverable D3.3. It has been revised and extended based on the experience collected in the project. For example, the dimension of a semantic content type has been added, since we learned in our work with the semantic desktop, that this is an important criteria to decide about preservation value.

In the following, we describe those dimensions together with examples for illustrating the concept of each dimension:

Content Type This dimension refers to the type of the content to be assessed. Types might be considered on several level ranging from image vs. text via specific formats towards more semantic types e.g. distinguishing a holiday photo from a food picture.

Investment In a wide sense, this dimension refers to the investment, which has been made into the resource and its improvement/change. For a photo collection such investment might be the annotation of photos, the improvement of photos in photo software or the creation of multiple photos of the same scene.

Gravity This dimension refers to the relationship or closeness of a resource to important events, processes, and structures in the domain under consideration. For personal photos this might be the closeness to an important event such as a wedding or an important life situations such as the first years of one's child.

Time Although the age of the content and time-related properties in more general are less important for long-term information management than for the decision about short term interests, temporal aspects still play an important role for assessing

preservation value. For social web content, for example, there is a trend to be more selective, when the content gets older.

Social Graph This dimension describes the relationship of the resource to the relevant social graph, i.e., the persons related to the resource, their roles and relationships. This might refer to the creators and editors of a resource as well as to persons related to the content of the resource.

Popularity This dimension refers to the usage and perception of the resource. For the case of social web content this might refer to shared and liked content.

Coverage & Diversity This dimension refers to factors, which considers the resource in relationship to other resources in the same collection. This includes factors such as diversity or coverage of sub-events, which are also used in making preservation decisions and, thus, influence preservation value. This can, for example, be taken into account by trying to cover all the sub-events of a holiday, when selecting a holiday photo collection.

Quality This dimension refers to the quality of the resource. Obvious examples for content quality is photo quality assessing, e.g., if the photo is blurred or exhibits good contrast. More advanced quality aspects are for example photo composition and aesthetics.

With respect to these dimensions, an interesting category of evidences is the one that signals redundancy. This might be for example a sequence of near-duplicate photos taken from the same scene or several versions of the same documents. Actually, redundancy can, on the one hand, be treated as a signal for engagement and investment (Investment Dimension): many photos are taken to ensure a good picture. On the other hand, redundancy is also a signal suggesting reduction (Coverage & Diversity Dimension): one would tend not to preserve many very similar photos.

For a better understanding of Preservation Value, we have investigated relevant factors within those dimensions for the case of determining preservation values for photos in a photo collection and for content in social networks. The results of those experiments are described in Section 3 and Section 5.2, respectively.

2.2 Appraisal

In general there is a common understanding that preservation also is a selection process [Lavoie and Dempsey, 2004], which introduces the activity of appraisal. Usually, the guiding principle for appraising records depends on the values of records, which can be categorized into two parts: 1) Primary value, and 2) Secondary value. The primary value refers to the value of records when they were created, and it can be further divided into three sub-values, i.e., administrative value, legal value, as well as financial value. The administrative value reflects the value of records for administrative functions in the organization. The legal value presents the value of records related to legal affairs, such as

materials (e.g., contracts) used for protecting legal right. The financial value indicates the value of records for the continuity of business, which includes evidence of financial transactions (e.g., budgets, invoices, etc). The secondary value is related to the enduring value for a society, like the value in areas of history, research, military, and so on.

In ForgetIT, we mainly focus on the secondary value and investigate methodological insights that could be used to automate the selection process. In particular, in this deliverable, we conducted studies on how to automatically estimating preservation value of records, including preservation value for images, text, as well as social media. More details about these works will be given in the next chapters.

In the conventional archive situation, a set of criteria provides guidance to decide which type of records should be selected for preservation to better satisfy the organization's purposes, such as what the organization was supposed to do, and then identify which records match those purposes predefined by a set of functions. A broad functions could be identified for an organization, like assuring organization institutional continuity, maintaining research and diffusing knowledge. Each function can be further divided into a set of subfunctions.

In the following, we will link the preservation value dimensions proposed above to appraisal methods and criteria used in archives. An obvious link can be made from the approach of functional appraisal and macro-appraisal [Cook, 2005] to the dimensions of **gravity** and the **social graph**. As in the case of **gravity**, macro-appraisal looks into the importance of information items for the institution, looking into the structure and processes of the institution. Functional appraisal looks at "the functions carried out by the record creator"[2] and uses this information in appraisal instead of content criteria, which is related to the dimension of the **social graph**, which has a broader scope enabling its application in personal and organizational settings. Another aspect, which is linked to the dimension of **gravity** is the aspect of long-term historical importance, which is often used as a criteria for assessing secondary value in appraisal. For dimension of **investment**, our project refers to the preparation of resource and the corresponding improvement. However, previous archivists would take the legal risks or administration efforts into account. The dimension of **popularity** is linked to the idea of *Social Significance* as named as appraisal criteria in [Harvey, 2007]. Popularity can be seen as the measurable part of social significance, but does not fully cover this aspect.

The dimension **coverage & diversity** is not directly related to in appraisal criteria, but is implicitly linked to by a Macro-appraisal approach, which aim to cover the big picture by the appraisal strategy.

The dimension **time** is important in both previous archive situation and ForgetIT. Examples are the stress for a need of a re-appraisal process [Conway, 2000] and the general understanding that in the digital age preservation decisions have to be taken in a timely fashion [Harvey, 2007].

For the dimension **content type**, we stress on the semantic type of a content object rather

[2] http://www.paradigm.ac.uk/workbook/appraisal/appraisal-approaches.html

than on its format as a decision criteria for preservation. This is again linked to a functional approach to appraisal.

Finally, the dimension **quality** is mainly relevant, when there are alternatives of different qualities to be selected from (e.g. redundant content). This dimension is clearly also relevant for archival appraisal settings, and in both cases a secondary criteria.

3 Preservation Value for Images

We continued investigating the problem of estimating the Preservation Value for images, particularly for personal photos. We chose to dedicate further effort to this scenario because nowadays people are getting more and more sensible to the problem of managing their own personal collections. As a matter of fact, photo taking is effortless, tolerated nearly everywhere, and makes people easily ending up with hundreds of photos taken during one single event (e.g a holiday trip). Simply dumping photos on some cheap storage device does not only introduce the risk of losing photos due to "digital forgetting", but it also often ends up with having "dark archives" of photo collections, which are rarely accessed and enjoyed again due to the great effort and time to be spent for revisiting, sorting, annotating.

In this deliverable, we extend the work previously reported in [Kanhabua et al., 2015] under the following aspects: (i) we involved more users and acquired more photo collections for our experiments (91 collections from 42 users, more than 18,000 photos in total); (ii) we investigated the role of coverage in personal photo selection and compared with state of the art methods; (iii) we performed an extensive evaluation and comparison of the different considered methods; (iv) under the assumption that the photo selection task can exhibit some degree of subjectivity, we experimented how to develop personalized selection models; (v) we included further and to some extent orthogonal information into the selection model, such as aesthetics, face clustering, sentiments.

3.1 Personal Photo Selection

We present our approach to select photos for preservation and revisiting, which has been published at ICMR '15 [Ceroni et al., 2015b]. It determines the preservation value for personal photos, with the goal of identifying those photos that are most important to the user to invest more effort in keeping them accessible and enjoyable.

Let the photo collection P be a set of N photos, where $P = \{p_1, p_2, \ldots, p_N\}$. The photo selection problem is to select a subset S of size θ ($S \subset P$ and $|S| = \theta$), which is as close as possible to the subset S* that the user would select as the photos most important to her, i.e. S meets user expectations.

Given a photo collection, we extract information from the images by applying different image processing techniques developed in WP4. Our main approach is named Expectation-oriented selection (Section 3.1.1), which learns to generate selections by taking into account user selection from personal collections as training data. Furthermore, we present two different Hybrid Selection methods (Coverage-driven, Optimization-driven), with the goal of investigating whether our method can be improved by combining it with state-of-the-art methods that explicitly consider coverage. The Hybrid Selection methods will be discussed in detail in Section 3.1.2.

3.1.1 Expectation-oriented Selection

Current approaches to photo selection for summarization aim at creating summaries that resemble the original collection as much as possible, [Li et al., 2003, Rabbath et al., 2011, Seah et al., 2014, Sinha et al., 2011]. We claim that selecting photos that are important to a user from personal collections is a different task than generating comprehensive summaries: the set of images important to the user might not be a proportioned subsample of the original collection. For instance, a user might ignore photos depicting joyless or boring moments. For this reason, we do not impose a strict notion of coverage but rather consider clusters and other global information as a set of features, along with photo-level features, learning their different impact in a single selection model. Our method does not require any manual annotation (e.g. tags, textual descriptions, file names) or external knowledge, differently from other works [Rabbath et al., 2011, Seah et al., 2014, Sinha et al., 2011].

The features are combined via machine learning, providing a model that predicts the probability of a photo to be selected, i.e. its importance. The selected sub-collection is created by ranking photos in the collection based on their predicted importance and by taking the top-k of them, where k is an input parameter and can assume any value lower than the collection size.

Features

Four groups of features, described below, have been designed to be used in the photo selection task, based on the information extracted from images via the image processing techniques developed within WP4. Please refer to [Papadopoulou et al., 2014] and [Solachidis et al., 2015] for a detailed description. In the following sections we will refer to the class of features using the names introduced hereafter, although the link between them and the preservation value dimensions defined in [Kanhabua et al., 2015] and in Section 2.1 is made explicit in their descriptions.

Quality-based features. They consist of the 5 quality measures described before: blur, contrast, darkness, noise, and their fused value. The assumption behind using this information is that users might tend to select good quality photos, although their impact seems to be less important in subjective selections of humans [Walber et al., 2014]. This family of features corresponds to the *quality* dimension defined in Section 2.1.

Face-based features. The presence and position of faces might be an indicator of importance and might influence the selection. We capture this by considering, for each photo, the number of faces within it as well as their positions and relative sizes. Each photo is divided in nine quadrants, and the number of faces and their size in each quadrant are computed. These features can be related to the *social graph* dimension defined in Section 2.1, because the presence of people in a photo can indicate relationships between the appearing people and the owner of the photo.

Concept-based features. The semantic content of photos, which we model in terms of concepts appearing in them, is expected to be a better indicator than low-level image

features, because it is closer to what a picture encapsulates. We associate to each photo a vector of 346 elements, one for each concept, where the i-th value represents the probability for the i-th concept to appear in the photo. The correspondence between this class of features and the Preservation Value dimensions is not strict and depends on what concepts are included in the concept space. Concepts might be related to *gravity*, in case they represent aspects related to the events in the collection, or to the *social graph*, in case they represent appearance of people, groups, or crowds.

Collection-based features. This family of features is a representative of the *coverage* dimension defined in Section 2.1. When users have to identify a subset of important photos, instead of just making decisions for each photo separately, the characteristics of the collection or a cluster a photo belongs to might influence the overall selection of the subset. For each photo, we consider the following collection-base features to describe the collection and cluster the photo belongs to: size of the collection, number of the clusters in the collection, number of near–duplicate sets in the collection, size of the near–duplicate sets (avg, std, max, min), quality of the collection (avg, std), faces in the collection (avg, std, max, min), size of the cluster (avg, std, max, min), quality of the cluster (avg, std, max, min), faces in the cluster (avg). Since the redundancy introduced by shooting many pictures of the same scene can be evidence of its importance for the user, we also consider whether photos have near-duplicates or not, as well as how big is the near-duplicate set the photo belongs to. Shooting many similar pictures of the same scene can be regarded as a form of *investment*, because the user puts effort in replicating a scene to ensure its availability and quality.

Importance Prediction and Ranking

Given a set of photos p_i, their vectors $\boldsymbol{f}_{p_i}$ containing the features presented above, and their selection labels l_{p_i} (i.e. *selected* or *not selected*) available for training, a prediction model represented by a Support Vector Machine (SVM) [Cortes and Vapnik, 1995] is trained to predict the selection probabilities of new unseen photos, i.e. their importance. For new unseen collections, feature vectors $\boldsymbol{f}_p$ are constructed based on the information extracted from the images and the importance of each unseen photo p is computed as $I_p = M\left(\boldsymbol{f}_p\right)$, which is the probability of the photo to be selected by the user. Once the importance of each photo in the collection is predicted, the photos are ranked based on this value and the top-k is finally selected (with k being an input parameters).

3.1.2 Hybrid Selection

Given the wide exploitation of the concept of coverage by many state of the art methods, we want to better understand its role in photo selection, in order to see if and in which way our method can be improved by combining it with explicit consideration of coverage. Another motivation is that coverage resulted to be a highly considered factor from our previous user study [Ceroni et al., 2015a]. Therefore, we propose and investigate two ways of combining our importance prediction model with coverage-oriented photo selec-

tion methods, denoted *hybrid selection* methods and described hereafter. Although kept into account within the expectation-oriented selection via the *collection-based* features (Section 3.1.1), the *coverage* dimension (Section 2.1) is dominant and explicitly considered in this family of selection methods. The *diversity* dimension is explicitly considered in the hybrid method described in Section 3.1.2.

Coverage–driven Selection

The coverage-driven selection is based on the widely used two-step process of first clustering and subsequently picking photos from the clusters. First, for a given collection C, a set of clusters CL_C is computed using the clustering techniques developed in WP4 and the importance $I(p)$ of each photo $p \in P_C$ is computed according to our importance prediction model (Section 3.1.1). Given the clusters CL_C, we use the importance $I(p)$ for each photo $p \in P_C$ to pick an equal number of top-ranked photos from each cluster in order to produce the selection S of required size k.

Cluster Visiting. When picking photos from each cluster, there are different possible ways of iterating over them until the requested size of the selection is reached. We experimented a round-robin strategy with a greedy selection at each round. Given an initial set of candidate clusters CL_{cand}, the greedy strategy in each step selects the cluster cl^* containing the photo p^* with the highest importance, according to the prediction model M. The photo p^* is added to the selection S and removed from its cluster cl^*. The cluster cl^* is then removed from the set of candidate clusters for this iteration, and the greedy strategy is repeated until the candidate set is empty. Once it is, all the not empty clusters are considered available again and a new iteration of the cluster visiting starts. This procedure continues until the requested selection size k is reached.

Optimization–driven Selection

Sinha et al. [Sinha et al., 2011] modeled coverage as part of a multi-goal optimization problem to generate representatives summaries from personal photo collections that resemble the original collection as much as possible. In more detail, in this work *quality*, *coverage*, and *diversity* of the summary are jointly optimized and the optimal summary S^* of a requested size k is defined as $S^* = \arg\max_{S \subset P_C} F\left(Qual\left(S\right), Div\left(S\right), Cov\left(S, P_C\right)\right)$, where $Qual\left(S\right)$ determines the interestingness of the summary S and it aggregates the *interest* values of the individual photos in the summary, $Div\left(S\right)$ is an aggregated measure of the diversity of the summary measured as $Div\left(S\right) = \min_{p_i, p_j \in S, i \neq j} Dist\left(p_i, p_j\right)$, and $Cov\left(S, P_C\right)$ denotes the number of photos in the original collection C that are represented by the photos in the summary S in a concept space.

We incorporate our expectation-oriented selection within this framework, creating the *optimization–driven selection*, by computing the $Qual\left(\cdot\right)$ function in the cost functional based on the importance prediction model (Section 3.1.1), that is $Qual\left(S\right) = \sum_{p \in S} M\left(p\right)$. Please refer to [Ceroni et al., 2015b] for further details.

3.1.3 Experiments

Experimental Setup

Dataset. We repeated the user study described in [Kanhabua et al., 2015] with more participants, which were asked to provide their personal photo collections and to select the 20% that they perceive as the most important for revisiting or preservation purposes. We obtained 91 collections from 42 users, resulting in 18,147 photos. The collection sizes range between 100 and 625 photos, with an average of 199.4 (SD = 101.03).

Evaluation Metrics. We evaluate the different methods considering the precision P@k of the selection S of size k that they produce, computed as the ratio between number of photos in S that were originally selected by the user and the size of S. The size k is considered as a percentage of the collection size. Statistical significance, performed using a two-tailed paired t-test, is marked as $^{\blacktriangle}$ and $^{\triangle}$ for a significant improvement ($p < 0.01$ and $p < 0.05$, respectively), and significant decrease with $^{\blacktriangledown}$ and $^{\triangledown}$ ($p < 0.01$ and $p < 0.05$, respectively) with respect to the baselines.

Parameter Settings. The classifiers employed for importance prediction and cluster filtering, built using the Support Vector Machine implementation of LibSVM, have Gaussian Kernels ($C = 1.0$, $\gamma = 1.0$) and have been trained via 10–fold cross validation.

Baselines

Clustering. For a given collection C, a set of clusters CL_C is computed. The selection is built by iterating the clusters, temporally sorted, in a round–robin fashion and picking at each round the most important photo from the current cluster (until the requested selection size is reached). The importance of each photo $p \in P_C$ is modeled as $I(p) = \alpha \cdot \left\|\boldsymbol{q}_p\right\| + (1-\alpha) \cdot \dim(F_p)$, which is a weighted sum of the quality vector of the photo and the number of faces in it. We experimented with different values of the parameter α, identifying the best value as $\alpha = 0.3$, which gives more importance to the number of faces in the photos. We report the performances obtained with this parameter value in our evaluation.

Summary Optimization. We implemented the approach presented in [Sinha et al., 2011] as another baseline, where summaries are generated by optimizing *quality*, *coverage*, and *diversity* as in Section 3.1.2. The *quality* of summaries is computed by summing the *interest* of photos in it, defined as a measure depending on photo quality and presence of portraits, groups, and panoramas. We computed the interest of photos as in the original work, using the concepts *face*, *3 or more people*, and *landscape* available in our concept set to represent portraits, groups, and panoramas respectively. Also *diversity* and *coverage* of summaries are computed coherently with their original computation, as already described in 3.1.2. Giving equal weights to the α, β, γ parameters gave us the best results, thus we will report the performances for only this setup in the following evaluation, denoting it *SummOpt*.

	P@5%	P@10%	P@15%	P@20%
Baselines				
Clustering	0.3741	0.3600	0.3436	0.3358
SummOpt	0.3858	0.3843	0.3687	0.3478
Expectation-oriented Selection				
quality	0.3431	0.3261	0.3204	0.3168
faces	0.4506▲	0.3968▲	0.3836△	0.3747△
concepts	0.5464▲	0.4599▲	0.4257▲	0.4117▲
photo-level	0.5482▲	0.4760▲	0.4434▲	0.4266▲
all (Expo)	0.7124▲	0.5500▲	0.4895▲	0.4652▲

Table 1: Precision of the expectation-oriented selection, for different sets of features.

Results

Expectation-oriented Selection. We evaluated our expectation-oriented selection with respect to the two baselines defined in Section 3.1.3. Different importance prediction models have been trained by using the subsets of the features described in Section 3.1.1. Since each group of features is linked to part of the preservation value dimensions (Section 2.1), our analysis provides insights about the importance of the dimensions in the context of personal photo selection for preservation. The results for different selection sizes (k) are listed in Table 1. The two baselines exhibit comparable performances, with *SummOpt* performing slightly better for all considered values of k (5%, 10%, 15%, 20%).

The *quality* features are the ones that perform weakest individually, which has already been observed for other photo selection tasks [Walber et al., 2014]. This corroborates the idea that low quality photos might be kept anyway because they contain and recall memories and event important to the user. *Faces* features alone already show better performances than the baselines. The performance achieved when only using *concepts* features is better than the ones of *quality* and *faces*: they are able to capture the semantic content of the photos, going beyond their superficial aesthetic and quality. The model trained with the combination of all aforementioned features, denoted *photo-level* because the features are extracted from photo level, slightly improves the performance of using concept features alone. This indicates that leveraging quality and faces features in addition to semantic measures, such as concepts, can better the overall performance.

If we include global features for each photo representing information about the collection, the cluster, and the near–duplicate set the photo belongs to, we get a comprehensive set of features, which we call *all*. The precision of the selection for this global model further increases for every selection size: this reveals that decisions for single photos are not taken in isolation but they are also driven by considering general characteristics of the collection the photo belongs to: e.g. number of photos, clusters, average quality of photos in the collection and in the same cluster, how many duplicates for the photo there are. This is a point of distinction with respect to state-of-the-art methods (represented by the

Info Gain	Feature Name	Info Gain	Feature Name
0.10836	ND of photos	0.01561	Avg aggr. quality in collection
0.02569	Images without ND in collection	0.01538	Std ND set size
0.02258	Min darkness in cluster †	0.01523	Min ND set size
0.02251	Std aggr. quality in collection	0.01469	Std faces in collection
0.02240	Norm of concepts in collection	0.01440	Concept “person”
0.02189	Count of faces in photo	0.01414	Count of faces in cluster†
0.02177	Avg size of ND sets in collection	0.01321	Std aggr. quality in cluster†
0.02144	Avg contrast in cluster†	0.01306	Concept “dresses”
0.02009	Max cluster size in collection	0.01291	Concept “joy”
0.01863	Avg contrast in collection	0.01273	Avg blur in cluster†
0.01760	Count of central faces in photo	0.01147	Avg blur in collection
0.01732	Avg count of faces in collection	0.00952	Concept “two people”
0.01610	Min clusters size	0.00889	Concept “entertainment”
0.01609	ND sets in collection	0.00873	Contrast of photo
0.01565	Size of central faces in photo	0.00826	Concept “girl”

Table 2: Top-30 features ranked by Information Gain with respect to the class.

two baselines), because our selection approach does not strictly handle collection-level information by imposing clustering (*Clustering*) or optimizing measures like coverage and diversity along with photo importance only based on quality and presence of people (*SummOpt*). It rather takes this global information in consideration in a flexible way through a set of features, whose impact to the selection is learned from user selections and expectations.

Feature Analysis. For sake of completeness, in Table 2 we report the top–30 features ranked based on the Information Gain with respect to the class (i.e. user selections). Despite the presence of similar and redundant features, the table still provides an overview of the features that are correlated to the class the most. The symbol † for features related to clusters means that the cluster containing the input photo is considered. For instance, given an input photo, the feature *Min darkness in cluster* represents the minimum darkness over all the images within the cluster the input photo belongs to. The first-ranked feature, whose Information Gain value is significantly higher than the ones of the other features, represents the number of near-duplicates that the input photo has. This reveals that the redundancy introduced by taking many shoots of the same scene is a strong signal of importance for that scene. Besides this feature, the other ones in the table have much smaller and similar Information Gain values. Many other high-ranked features are computed considering global information from clusters and collections. Features computed based on faces are also important. Quality is mostly considered in relation to collections and clusters (i.e. quality statistics with respect to the whole collection or a given cluster). A relatively low number of features represent concepts, which is somewhat counter intuitive if compared with the selection results of the *concepts* features reported in Table 1. Nevertheless, the high performance values, if compared to those of *quality* and *faces* features, might be due to the combination of many concept features, although they are not all top-ranked.

Expectation vs. Hybrid Analysis. We now compare the expectation-oriented selection model exploiting all the available features (*Expo*), and the hybrid selection models. The results of the Hybrid Selection methods are listed in Table 3, where they have been split

	P@5%	P@10%	P@15%	P@20%
Baselines				
Clustering	0.3741	0.3600	0.3436	0.3358
SummOpt	0.3858	0.3843	0.3687	0.3478
Coverage-driven Selection				
basic	0.4732▲	0.4113▲	0.3902△	0.3809△
greedy	0.6271▲	0.4835▲	0.4391▲	0.4262▲
SummOpt++	0.7115▲	0.5533▲	0.4937▲	0.4708▲
Expo	0.7124▲	0.5500▲	0.4895▲	0.4652▲

Table 3: Precision of the hybrid selection methods.

based on the two different classes of hybrid selection. For coverage-driven selection, we report results of different combinations: *basic* refers to the coverage–driven selection which only uses our importance prediction model defined in Section 3.1.1 as photo importance measure, picking photos in a round-robin fashion from clusters temporally ordered; *greedy* indicates the use of the greedy visiting strategy. The optimization-driven method is referred to as *SummOpt++*.

Considering Table 3, we can observe that the performances of *Expo* are better or comparable with the ones of the hybrid-selection models. In particular, the improvements of *Expo* with respect to the *coverage-driven* methods are statistically significant. The only improvements over *Expo* (which anyway are not statistically significant) are obtained when considering methods that possess a relaxed consideration of coverage and global information in general (*SummOpt++*). These results further support our assumption that in our photo selection task a strong consideration of coverage overstresses this aspect as a selection criterion. Only for the methods with a more flexible consideration of coverage the performances are similar to the pure expectation-oriented method.

Features and Preservation Value Dimensions. This last part summarizes the main insights obtained from this work, linking the results of photo selection to the high-level dimensions of preservation value (Section 2.1). From the results reported in Section 3.1.3, the *quality* dimension seems not to be of primary importance for preservation in personal scenarios. As an example, one might want to keep a photo because it evokes memories of the time when we took the photo, despite its low quality. The *faces* class of features alone also was not a very good indicator . The introduction of more powerful and demanding processing techniques like face clustering and tagging might probably help make the *social graph* dimension more important (at the prices of increasing the investment of the user in tagging and annotating).

The high expectations on the *coverage* dimension were not confirmed by the experimental results, since we observed that emphasizing coverage did not yield to significant improvements over the pure expectation-oriented selection. The only positive result related to coverage is the high correlation between the presence of near-duplicates and selec-

tion decisions (Table 2), which shows that people tend to shoot many similar pictures of what they like the most and is most important to them. However, this fact is more related to the concepts of redundancy and investment than coverage. In our opinion, one of the main pitfalls of stressing coverage to emulate human selections from personal collections for preservation is that not all the clusters are usually equally important for the users. The optimal parameter values identified for the optimization-driven selection (Section 3.1.2), jointly considering importance, coverage, and diversity, showed that also the *diversity* dimension had a low impact in the selection. While being widely considered for photo summarization, diversity resulted to have only a marginal role in emulating user selections for preservation.

3.2 Personalization

The selection method described in Section 3.1 generates one single selection model to be used for any user and input collection. As a matter of facts, the photo selection process (especially for personal data) can be highly subjective and the factors that drive the selection can vary from individual to individual [Savakis et al., 2000]. Some users might be particularly interested on photos depicting many people, while others might prefer pictures with landscapes or buildings. Therefore we investigated how to develop personalized photo selection models to assist users in photo selection, which adapts to the photo selection behaviors and preferences of the user. Starting from the general model presented in Section 3.1, selection decisions done by a given user on new collections are acquired and the selection model is updated according to them. Feeding the revisions of the user for automatically generated selection back into the selection model can, on the long run, bridge the gap between the general selection model and the user preferences. Moreover, in order to tackle the problem of having limited initial data to train the model (cold-start scenario), we experiment whether the exploitation of data from other users can boost the adaptation of the model to a given user when a limited amount of personal training data is available.

Previous works on photo selection [Obrador et al., 2010, Yeh et al., 2010] have revealed that the photo selection task is, to some extent, subject to the preferences of each user. General selection models, although capable of representing common selection patterns (e.g., photos depicting people might be usually appreciated), might be improved by considering the preferences of each single user separately and derive personalized models for them. In this section, we show how personalized models have been derived from the photo selection approach described in Section 3.1, denoted *general model* hereafter.

We adopt an incremental learning strategy to achieve personalization, re-training the model each time new data (i.e. selection decisions) is provided by the user. The annotated photo collections available to train the general model are first pre-processed through image processing techniques and features are extracted from them, in the same way described in Section 3.1. For each new collection provided by the user, a first selection is made by the trained general model and the selected photos are displayed to the user, who gives feedback revising the automatically generated selection. The training dataset

is then expanded by adding the feedback data and the general model is retrained with the updated training dataset. Iterating this process, it is expected that the gap between user expectations and model's selections gets lower, due to the adaptation of the model towards the selection preferences of the user.

Incremental Learning

A recurrent problem in machine learning is continuously managing new data, so that the existing model can be updated to accommodate new information and to adapt to it. Two common approaches for updating the model to new incoming data are *online learning* [Bordes et al., 2005], where the model is updated only considering the new data, and *incremental learning* [Cauwenberghs and Poggio, 2001], where the model update considers the old training data along with the incoming data. We consider the latter strategy because, in our scenario, the updated model has to be aware of the entire data available, not just of the most recent one.

Although efficient and effective incremental versions of off-line learning algorithms exist (e.g., [Cauwenberghs and Poggio, 2001]), we perform the model update by including the new data in the training set and re-train the model from scratch. We implemented such more straightforward but functionally equivalent approach because our scenario does not impose strict time constraints for the model update, thus making the efficiency benefit of incremental versions of secondary importance. The time taken by a user to produce a new collection (e.g. after a trip or vacation) can be considered sufficient to re-train the model with the whole available data. Should the temporal constraints of the envisioned scenario become stricter, the incremental version of the employed algorithm could be plugged in without changing the functionalities of the whole application.

Model Update

Our personalized photo selection models, one for each given user, are built by re-training the model every time that a new collection is imported and the automatic selection done by the current selection model is revised by the user. The procedure of the model update is the following. The input includes a set of new unseen collections $C = \{C_1, \ldots, C_n\}$ from the user as well as a set of collections C^* with selection labels available, which represents the initially available training data. The output is the set of the test collections with prediction labels (selected or not selected) which is denoted as $C' = \{C'_1, \ldots, C'_n\}$. At the beginning, the training dataset T is composed by the initial data C^* and an initial prediction model M is trained from it applying the method described in [Ceroni et al., 2015b]. For each photo p in the user collection C_i, the selection probability (i.e. importance) i_p is predicted by the general model M and added in the importance list which records the importance of photo in the entire collection. Following, according to [Ceroni et al., 2015b], the photos are ranked based on their importance value and top-n of them are selected which results in the selections C'_i. In order to know which photos the user would really have selected or not selected, we ask the user to give feedback by revising the generated

selections. This is finally included within the available training dataset. The prediction model M will be retrained by using such new training data and applied to make predictions for the next coming collection C_{i+1} of the user.

Cold–Start Problem

Usually, the adaptation of a system within the initial rounds of user interactions is affected by the so called *cold-start problem*: there is not enough (or even not at all) training data to let the model adapt to the user. This holds in our scenario as well, where the selection model might not perform proper predictions because of the lack of annotated collections in the initial training set T. We consider two ways of building the initial training set. One consists in using one annotated collection of the given user as initial training set. The other is based on using annotated collections from other users to train the initial selection model, based on the assumption that some common selection patterns could be captured through a sample of selections done by other users. We will experiment and compare these two strategies in our experiments.

3.2.1 Experiments And Results

Experimental Setup

Dataset. We used the same dataset described in Section 3.1.3 for our experiments. In order to assess personalization performances, we consider users who contributed at least 5 collections as test users. Among the overall 91 photo collections, there are 11 users who provided more than 5 collections (10 users contributed 5 collections, 1 user contributed 6 collections) which result in 56 collections totally. According to this, our dataset is split into two parts: one part contains 35 collections from 31 users, whereby each user provided at most 2 collections, which is called *general dataset*; another part contains 56 collections from 11 users, whereby each user provided at least 5 collections, which is called *personalized dataset*.

Evaluation Metrics. The evaluation metrics are the same as the ones reported in Section 3.1.3. In particular, we compute the precision for $n = 20\%$, which is indicated as P@20%, coherently with our user study where participants were asked to select the 20% most important photos from their collections. In order to assess the adaptation of our personalized model to users, we apply the personalization process described at the beginning of Section 3.2 to the collections of each user separately and average the P@20% among the test collections available at each iteration k, where k denotes the number of collections that are used for training the personalized model.

Parameter Settings. The classifier employed for importance prediction, built using the Support Vector Machine implementation of LibSVM[3], has Gaussian Kernels and has been trained via 10–fold cross validation on the training set. Note that the training set is ex-

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

panded at each iteration (i.e. each time a new annotated collection of the user is provided), and the training via 10–fold cross validation is repeated each time. The open parameters were tuned via grid search and updated at each iteration. The ones identified for the *general dataset* where $C = 1.5$, $\gamma = 0.25$.

Training and Test Sets

We evaluate the performances of the model update (Section 3.2) over different rounds of adaptation. The *personalized dataset* is split by users where each user owns 5 collections (one user owns 6). At each iteration k, for each user with N collections, k collections are added to the initial training set to learn the personalized model of the user, and $N - k$ collections are used for testing. The ways in which the original training set is built are described in Section 3.2.1.

We experiment all the values of k ($k = 0, 1, 2, 3, 4$), and for each of them we repeat the split and evaluation 5 times so that all the collections could be selected the same times as training collections. Note that the iteration $k = 0$ corresponds to the situation when the selection model is trained only on the initial training set. The selection strategy to select training collections is the following. When $k = 1$, we ensure that each collection of the user that we are considering is selected once as initial training data and the remaining four collections are treated as test data, then we average the performances. When $k = 2$, we pick two collections at each time from 5 collections, with the constraint that each collection could only be selected twice in all 5 repetitions (to be fair to all collections). We then average the performance achieved at each time. The cases when $k = 3$ and $k = 4$ can be done in the same manner. Finally, we average the performances over users for the same value of k.

Different Training Sets

The three considered ways of building training sets are described hereafter. The model update and the split in train and test set previously described are the same in each case.

Stand-alone. The initial model is trained with one random collection of the user, and the model update is incrementally done considering the remaining collections (starting from iteration k = 1). The iteration $k = 0$ is not considered since the training set would be empty at this stage.

Collaborative. The initial training set at $k = 0$ is formed by all the collections within the *general dataset*. This case represents the situation where, in absence of large amount of annotated personal data for training, annotated collections of other users are used to alleviate the cold-start problem.

User-agnostic. Similarly to the *collaborative* case, the *general dataset* is used as initial training set. However, at each iteration k, instead of including k collections of the user that we are considering, we add k randomly selected collections from the other test users.

	k = 0		k = 1		k = 2		k = 3		k = 4	
	P@20%	Δ	P@20%	Δ	P@20%	Δ	P@20%	Δ	P@20%	Δ
Stand-alone	-	-	0.353 ± 0.060	-	0.374 ± 0.068	+5.9%	0.383 ± 0.067	+2.4%	0.402 ± 0.069	+5.0%
Collaborative	0.427 ± 0.057	-	0.430 ± 0.054	+0.7%	0.432 ± 0.055	+0.5%	0.437 ± 0.050	+1.2%	0.444 ± 0.061	+1.6%
User-agnostic	0.427 ± 0.057	-	0.427 ± 0.052	+0.0%	0.428 ± 0.055	+0.2%	0.429 ± 0.053	+0.2%	0.426 ± 0.055	-0.7%

Table 4: P@20%, standard deviation, and performance gain of the personalized models at each iteration.

This case is motivated by the assumption that, if one collection, which is not from the user that we are considering, is included in the training set at each iteration, then the adaptation performances should be smaller than including collections that are from the user that we are considering. This would highlight the importance of incorporating selection information of the user in the training set when making selections for new collections of the same user.

Results

As a motivation to the need of personalization in photo selection, we trained a not personalized selection model on the *general dataset* and we tested its performances (P@20%) on the *personalized dataset*. Looking at the results, we observed a large amount of variability in performances over the different collections, with precision values ranging between 0.190 and 0.722. The same pattern was observed when grouping collections by test users, although the differences in performances were less prominent. This shows that a single selection model has limitations in meeting the expectations and preferences of different users, and the overall performances of the system could be improved by learning selection models personalized to each single user.

The results of our personalization procedure, considering the three different ways of constructing the training set described before, are shown in Table 4. Along with the precision when selecting the 20% of the original collection (P@20%) and its standard deviation over the test users, we also explicitly report the relative gain (Δ) obtained between two consecutive iterations. For instance, the Δ for $k = 3$ represents the relative gain in P@20% with respect to the one achieved for $k = 2$. It is possible to observe that the precision of both *stand-alone* and *collaborative* increases at each iteration, i.e. with the increase of the number of user's collections considered for training the model. This shows that having a selection model partially aware of the user preferences (by exploiting a certain amount of the selection behavior in the training phase) can improve the precision of new unseen collections of the same user. The precision of *collaborative* is higher than the one of *stand-alone*, especially at the first iterations, showing that the selection data from other users can alleviate the cold-start problem. The gain Δ of *stand-alone* at each iteration is higher than the one of *collaborative*, because the initial model is weaker (due to the limited training set) and the inclusion of new training collections has a higher impact on the learning. It is important to clarify that the standard deviation observed in these experiments is relatively high. This can be due to a mixture of aspects, such as (i) a limited size of test set (both in terms of users and iterations), (ii) intrinsic changes of difficulty among

collections of the same user. For this reason, although a promising adaptation to the user emerges from our results, the inclusion of a wider data set would be required to show it more significantly.

Comparing *user-agnostic* and *collaborative*, the former exhibits an almost null gain in performances over iterations (it is even negative for $k = 4$), while the latter leads to a higher and increasing performance gain iteration after iteration. This shows that the increase of performance at each iteration is due to the inclusion of a new collection of the same user in the training set and not simply caused by the fact that the training set is expanded at each iteration, since in this case the gain of *user-agnostic* should have been higher as well. Given the relatively high values of standard deviation, this promising result would require an extended number of test collections and iterations to be more evident and statistically significant.

3.3 Exploiting Additional Information

Working on top of the selection model described in Section 3.1, we included additional information automatically extracted from images to make the model richer and capable of generating more precise selections. The goal is finding useful information that can be used to model those selection patters that are still hidden and not considered in the previous selection model. Such information is translated into different sets of features, which are added to the ones already available and exploited during the learning process.

In the next sections we will describe the different extracted features and show the results that we achieved when including them in the learning process.

3.3.1 Feature Description

Given an input image, we extracted different types of information, which are to some extent orthogonal to each other and together can give a more comprehensive description of the image's content. This information consists in image aesthetics, low-level content information, emotions, and face clustering.

Low-level Information

We have implemented part of the features presented in [Machajdik and Hanbury, 2010], where the authors investigated how to leverage low-level content information to predict emotions and sentiments arising from pictures.
HSV Statistics. We represented pictures in the HSV color space and we computed statistics (avg, std, min, max) for Hue, Saturation, and Brightness.
Pleasure, Arousal, Dominance. A psychological experiment [Valdez and Mehrabian, 1994] showed that particular linear combinations of Saturation and Brightness fairly correlate with the sentiments of pleasure, arousal, and dominance. We then computed such lin-

ear combinations, which are $0.69Y + 0.22S$ for pleasure, $-0.31Y + 0.60S$ for arousal, and $0.76Y + 0.32S$ for dominance.

Colorfulness. We measured the colorfulness of an image by computing the Earth Movers Distance (EMD) between the histogram of an image and the histogram having a uniform color distribution (one for each R,G,B channel).

Color Names. Under the assumption that each color has a special meaning, we (i) used the algorithm presented in [van de Weijer et al., 2007] to classify pixels into one of the 11 basic colors (black, blue, brown, green, gray, orange, pink, purple, red, white, yellow), and (ii) counted the total number of pixels for each distinct color.

Textures. We computed Tamura texture features [Tamura et al., 1978], which, among others, can represent textural aspects like coarseness, contrast, directionality.

Dynamics. Studies (e.g. [Itten, 1973]) have suggested that the presence and slope of lines in pictures can trigger different emotions. For instance, horizontal lines are associated with calmness, while slant lines indicates dynamism. Therefore, we identified lines in images and counted the number and length statistics of static lines (horizontal and vertical) and slant lines (a line was classified as static if its angular coefficient was within $[-15; 15]$ or $[75; 105]$).

Skin. The amount of skin in an image is a signal of people appearance in images. Therefore, we considered the color spectrum suggested in [Liensberger et al., 2009] that represents the color of skin in the YCbCr color space, and we counted the percentage of pixel belonging to it.

Image Aesthetics

Image Aesthetics can reflect how an image is attractive and pleasant to the observers, for instance considering how colors, shapes, and objects are arranged in the image content. Along with the already considered image quality, aesthetics contributes to model the *quality* dimension defined in Section 2.1. We took inspiration by previous approaches in computational aesthetics [Yeh et al., 2010, Mavridaki and Mezaris, 2015] to derive the following aesthetics features. Some of these features have been provided by WP4.

Rule of Thirds. The *Rule of Third* is a well-known composition guideline, based on the idea of splitting the image content in vertical and horizontal thirds and placing the main subjects at their intersections (also called *power points*). First, the main subjects were identified by (i) segmenting the image and (ii) assigning a saliency score [Achanta et al., 2009] to each segment by averaging the saliency of pixel belonging to the segment. Second, the rule of third is measured by aggregating, for each segment, it's size, saliency, and distance to the closest power point. Intuitively, main subjects close to power points will make the feature value higher.

Simplicity. We computed two values to represent the simplicity of the photo's content. The first one is computed by building the Region of Interest (ROI) map based on saliency and then summing the sizes of all the not overlapping bounding boxes identified in the map. The second value, based on the idea that simplicity is the "attention distraction of the objects from the background" [Luo and Tang, 2008], has been calculated by (i) separating subject and background regions and (ii) using the color distribution of the background to

evaluate simplicity.
Contrast. We computed two measures of contrast, defined as the degree of diversity among the components of an image. The first one is the Weber Contrast, which assesses contrast in terms of the diversity of intensity values within the image. In order to consider color contrast, we also used the CIEDE2000 color difference equation presented in [Sharma et al., 2005].
Intensity Balance. Content balance can transmit equilibrium and calmness to who is watching the picture. We assessed balance in terms of pixel intensity, computing the difference between two intensity histograms, one for the left-hand and one for the right-hand part of the image.
Naturalness. We finally computed theI Naturalness index (CNI) defined in [Huang et al., 2004]. It is a value summarizing how natural the colours in an image are, where higher values indicates that the image colors are more natural.

Emotional Concepts

The concepts considered so far in the selection model have almost always a neutral meaning and interpretation. Concepts like animal, building, beach does not directly suggest any particular positive or negative sentiment. In order to introduce emotional and sentimental aspects in the photo selection, we applied the concept detectors available in SentiBank [Borth et al., 2013] to extract a set of 1200 Adjective Noun Pairs (ANP) from images. By definition, ANPs are formed by a noun, which represents a neutral concept, and an adjective, which instead associates a particular emotion to the concept. For instance, for the same neutral concept *cat*, the concept set contains its variants *sleepy cat*, *wet cat*, *lost cat*, *cute cat*, *playful cat*, *lazy cat*, *angry cat*, *grumpy cat*, etc. Each of these concepts, although always representing a cat, has a different emotional impact.

Face Clustering

Face detection, already considered in the selection model, is a signal of people appearance in photos. However, it does not reveal anything about the ”role” of a given face within the collection, for instance how much the person is popular in it (in terms of occurrence frequency). A person related to who took the photos, e.g. a friend, husband, wife, will probably occur many times in the collection. On the contrary, random people appearing by chance, e.g. in outdoor crowd environments, will have a low occurrence frequency. This information contributes to model the *social graph* dimension defined in Section 2.1, since it provides insights about the relationships between the people appearing in the pictures and the owner of the collection. The face clustering technique implemented within WP4 and described in [Solachidis et al., 2015] has been applied to model this. Each face cluster represents one distinct person and contains all the occurrences (faces) in the images within a collection. We leveraged this information to derive features about the popularity of faces and then to have aggregated measures of the popularity of an image. First, for each face, we compute its popularity as the size of the face cluster it belongs to

	P@5%	P@10%	P@15%	P@20%
Expo				
quality	0.3431	0.3261	0.3204	0.3168
faces	0.4506	0.3968	0.3836	0.3747
concepts	0.5464	0.4599	0.4257	0.4117
all	**0.7124**	**0.5500**	**0.4895**	**0.4652**
Expo++				
low level	0.4399	0.3913	0.3729	0.3697
aesthetics	0.4406	0.3923	0.3732	0.3639
face popularity	0.4692	0.4101	0.3977	0.3945
concepts (DCNN)	0.5694	0.4945	0.4553	0.4436
concepts (SentiBank)	0.6124	0.5172	0.4674	0.4502
all	**0.7426**$^{\triangle}$	**0.6155**$^{\blacktriangle}$	**0.5330**$^{\blacktriangle}$	**0.5121**$^{\blacktriangle}$

Table 5: Precision of the expectation-oriented selection enriched with additional feature sets.

(normalized by the total number of faces in the collection). Second, for each image, wec consider the popularity values of all the faces contained in it and compute statistics (avg, std, min, max) about them.

Concept Detection with Deep Learning

For sake of completeness, we mention that we extracted concepts values using a new version of concept detection developed in WP4 and reported in [Solachidis et al., 2016]. The concept set contains the same 346 concepts considered in the previous version, but the input features to train the concept detectors is different. Instead of using SIFT, SURF, and ORB local descriptors (and their color variants) for visual feature extraction, features learned via Deep Convolutional Neural Networks (DCNNs) are considered as input to the concept detectors. This set of features made the concept detectors considerably more accurate, and hopefully this will help in the task of photo selection as well. Please look at [Solachidis et al., 2016] for further details.

3.3.2 Results

Finally, we report the performances of the selection model when using the different previously described sets of features within the learning process. The experimental setup is the same one used for the evaluation of the original selection model (Section 3.1.3). The results are listed in Table 5, distinguishing over different subsets of features. The results referring to the experiments with the additional sets of features are under the name *Expo++*. We also report the results of the previous feature sets (Section 3.1.3) for sake of comparison.

The *Expo++* model exploiting *all* the additional features outperform the previous *Expo* model for all the selection sizes k, with relative improvements ranging from 11.9% (P@10%) to 4.2% (P@5%). The improvements have been proved to be statistically significant. This shows that expanding the selection model with a more variegate sets of features does help in improving the selection precision. Regarding the individual subsets of features, both *concepts (DCNN)* and *concepts (SentiBank)* improved the performances of the *concepts* features. This means that having both more precise concept detectors (*concepts (DCNN)*) and a set of concepts considering sentiments and emotions (*concepts (SentiBank)*) helps in the selection task. The inclusion of face clusters information to assess face popularity also exhibited a slight improvement over the *faces* features alone, although popularity features were expected to have a stronger impact. Both *low level* and *aesthetics* features resulted to be more useful than the mere *quality* features extracted via quality assessment, but still their performances are lower than the ones of the other features set (especially the ones related to concepts). This is a further confirmation that, for the task of photo selection from personal collections, the semantic and emotional aspects are dominant with respect to those related to surface visual content and aesthetics.

3.4 Integrating Multi-view Information

3.4.1 Multi-view Representation

Working on photos which are comprised of multiple views (or representations), for example, a photo can be represented by its visual contents, annotated tags, social comments, and so on. These different photo views usually provide complementary information to each other, and in this section we investigate how to integrate the multiple photo views effectively in order to obtain a better photo representation. In particular, we attempt to learn a new representation which can better reflect the underlying clustering structure of each view. The basic assumption, named *multi-manifold assumption*, is that the learnt representation should vary smoothly along the manifolds of different views, i.e., if two data points x_i and x_j are close in more view geometries, their corresponding coefficients s_i and s_j should be more close to each other with respect to the new basis B.

In the following sections, we will first introduce the objective function, and followed by the solutions for the optimization problem.

a) Objective Function.

To this end, we propose to exploit the manifold structure embedded in each view and incorporate them as set of graph Laplacian constraints into the sparse coding framework.

Formally, let $X^{(1)}, X^{(2)}, \cdots, X^{(n_v)}$ denote the n_v views. Here for the v-th view, we build a k-nearest neighbor graph, denoted as $G^{(v)}$, to encode its manifold information. Let $W^{(v)}$ be the weight matrix corresponding to $G^{(v)}$, where $w_{ij}^{(v)} = 1$ if x_i and x_j are among the k-nearst neighbors of each other with respect to the v-th view, otherwise $w_{ij}^{(v)} = 0$. We then define the Laplacian matrix as $L^{(v)} = W^{(v)} - D^{(v)}$, where $D^{(v)}$ is a diagonal matrix

with $(i,i)-$element equal to the sum of the i-th row of $W^{(v)}$.

In order to preserve the manifold structures of multiple views, we represent these manifold structures as a set of graph Laplacian constraints, which can be easily formalized as $\frac{1}{2}\sum_{i,j=1}^{n}||s_i - s_j||^2 W_{ij}^{(v)} = tr(SL^{(v)}S), v = 1, \cdots, n_v$, and incorporate these constraints into the objective function. Therefore, the objective function of MMRSC can be formalized as:

$$\begin{aligned} &\min_{B,S} ||X - BS||_F^2 + \sum_{v=1}^{n_v} \alpha_v Tr(SL^{(v)}S^T) + \beta \sum_{i=1}^{n} ||s_i||_1 \\ &s.t. ||b_i||^2 \leq c, i = 1, \cdots, m \end{aligned} \tag{3.1}$$

where X is the original data representation[4], n_v is the number of graph Laplacian constraints, and $\alpha_v \geq 0$ is the graph regularization parameter of the v-th manifold. When we increase α_v in Equation (3.1), the influence of the v-th manifold regularizer increases, and the corresponding effect is that s_i and s_j become more similar to each other if they are close in the v-th view. On the other hand, when we decrease α_v, the influence of the v-th manifold regularizer will decrease as well. In an extreme case, if we set all $\alpha_v = 0$, $v = 1, \cdots, n_v$, our approach will regress to the standard sparse coding.

The objective function in (3.1) is convex either in B or in S, while it is not convex in both of them simultaneously. For learning S and B, we resort to an iteratively optimization method as proposed in [Lee et al., 2007]. The optimization contains two steps: (1) fix the dictionary B while learning coefficients S; then (2) fix the coefficients S while learning the dictionary B. We iteratively execute these two steps until convergence, or until a pre-specified iteration number is reached.

b) Learning Sparse Coefficient Matrix.

In this section, we consider how to learn the sparse coefficient matrix S by fixing the dictionary B. For this purpose, the optimization problem (3.1) becomes:

$$\min_{S} ||X - BS||_F^2 + \sum_{v=1}^{n_v} \alpha_v Tr(SL^{(v)}S^T) + \beta \sum_{i=1}^{n} ||s_i||_1 \tag{3.2}$$

In order to facilitate manipulations in vector form, we rewrite the problem (3.2) as:

$$\begin{aligned} \min_{\{s_i\}} &\sum_{i=1}^{n} ||x_i - Bs_i||^2 + \sum_{i,j=1}^{n} (\sum_{v=1}^{n_v} \alpha_v L_{ij}^{(v)}) s_i^T s_j \\ &+ \beta \sum_{i=1}^{n} ||s_i||_1 \end{aligned} \tag{3.3}$$

Regarding the regularization terms $\sum_{i,j=1}^{n}(\sum_{v=1}^{n_v} \alpha_v L_{ij}^{(v)}) s_i^T s_j$ in the problem (3.3), each s_i is coupled with other coefficient vectors $\{s_j\}_{j \neq i}$. In order to solve this problem, we optimize

[4] In this paper, we leverage the concatenated representation as X. Note that other representations can also be considered as X.

over each s_i individually by keeping other coefficient vectors fixed, and get the following optimization problem for each s_i:

$$\min_{s_i} f(s_i) = ||x_i - Bs_i||^2 + (\sum_{v=1}^{n_v} \alpha_v L_{ii}^{(v)}) s_i^T s_i + s_i^T h_i + \beta \sum_{j=1}^{m} |s_i^{(j)}| \quad (3.4)$$

where $h_i = 2\sum_{j \neq i}(\sum_{v=1}^{n_v} \alpha_v L_{ij}^{(v)}) s_j$, and $s_i^{(j)}$ is the j-th coefficient of s_i.

Since problem (3.4) with ℓ_1-regularization is non-differentiable when s_i has values of 0, we cannot adopt the standard unconstrained optimization methods to solve this problem. Several approaches are available for solving this problem [Andrew and Gao, 2007, Lee et al., 2007, Schmidt et al., 2007]. In this paper, we follow an efficient solution proposed in [Lee et al., 2007], and use the feature-sign search algorithm to solve the problem (3.4).

c) Learning Dictionary.

For solving the optimization problem in (3.1) over the dictionary B, we fix the coefficients S and the problem reduces to a least squares problem with quadratic constraints:

$$\min_B ||X - BS||_F^2 \quad \text{s.t.} ||b_i||^2 \leq c, i = 1, \cdots, m. \quad (3.5)$$

There are several methods can be used for solving this optimization problem, in this paper, we choose the more efficient Lagrange dual method to solve the optimization problem [Lee et al., 2007]. Due to the limitations of space, here we only give the optimal solution for B as follows:

$$B = XS^T \cdot (SS^T + \Lambda)^{-1} \quad (3.6)$$

where $\Lambda = diag(\vec{\lambda})$, $\vec{\lambda} = [\lambda_1, \cdots, \lambda_m]^T$, and each $\lambda_i \geq 0$ is a dual variable. We refer the reader to [Lee et al., 2007] for more details.

3.4.2 Results

In this section, we empirically evaluate the proposed algorithm on a real-world photo dataset. The experimental results demonstrate the effectiveness of our proposed algorithm. The dataset used in our experiments is MirFlickr [Huiskes and Lew, 2008], which comprises 25,000 images from the Flickr[5]. We have two views of MirFlickr dataset, one is the 8,740 dimensional tag view and the other is the 305 dimensional visual view. For the tag view, we clean the raw tag data by removing stop words, converting letters into lower

[5]https://www.flickr.com/

case, and ignoring non-English tags. Moreover, we further discard tags with a frequency less than 3 and images with less than 2 tags in order to reduce the noise. Then we select 7,425 images from 10 categories, which are considered less correlated to each other. The number of images in each category varies from 100 to 1600 approximately. The tags are weighted by using the TF-IDF weighting scheme. While for the visual view, we use Lire [Lux and Chatzichristofis, 2008] to extract 305-D global features, including the 192-D Fuzzy Color and Texture Histogram [Chatzichristofis and Boutalis, 2008], 33-D MPEG-7 Color layout [Chang et al., 2001], and 80-D MPEG-7 Edge Histogram [Chang et al., 2001].

We compare our method with 7 baseline approaches: ConcatKmeans, ConcatNMF, ConcatSC, ConcatGraphSC, CollNMF [Akata et al., 2011], MultiNMF [Liu et al., 2013], and CoNMF [He et al., 2014]. The former 4 methods apply k-means, NMF, SC, and GraphSC over the concatenated data representation, respectively, while the remaining methods are the state-of-the-art work and attempt to learn a new representation of the data with different constraints. For a fair comparison of the different methods, we first apply all methods except ConcatKmeans to learn a new representation with the same dimension (e.g., a 64-dimensional vector) for the data, and then apply k-means algorithm on the new representation for clustering. Note that we can also learn a new representation with the same dimension as the number of ground-truth clusters where each dimension represents a cluster membership, and then select the maximal dimension as the final cluster label.

Due to the limitation of space, we only report the results of applying k-means on the learnt representation since it achieves better performance in our experiments. We carry out the experiments by conducting 20 test runs with different initializations. In MMRSC, the parameters β and k are empirically set as 0.1 and 3 respectively, and the parameters $\alpha_v(v = 1, \cdots, n_v)$ are uniformly set as 1. For simplicity, we use α instead of α_v $(v = 1, \cdots, n_v)$ for all views.

For evaluation, two standard clustering metrics, the accuracy (AC) and the normalized mutual information (NMI), are used to measure the performance.

d) Comparison.

As can be seen from Table 6, on the MirFlickr dataset, we find that the performance of ConcatNMF is better than that of ConcatKmeans. This shows that when the dataset is heterogeneous, directly applying the k-means clustering algorithm over a concatenated representation may not work effectively. Unsurprisingly, both ConcatSC and ConcatGraphSC are better than ConcatNMF and ConcatKmeans, due to the incorporation of the sparsity property. One interesting result is that ConcatGraphSC is worse than ConcatSC on MirFlickr, this is because the manifold structure based on the combined view is unreliable. The performance of CollNMF is comparable to that of ConcatNMF. This is consistent with the analysis that CollNMF is equivalent to conducting NMF on a combined view [Liu et al., 2013]. The performance of MultiNMF is worse than ConcatKmeans because MultiNMF can perform well only when the dataset is homogeneous [He et al., 2014]. Regarding the CoNMF method, it is interesting to see that the performance of CoNMF-W and CoNMF-B vary greatly. CoNMF-B outperforms all other baseline methods, reach an accuracy of 0.366, while CoNMF-W underperforms all other baseline methods, with an

Table 6: Clustering performance (mean $\pm$ standard deviation) on the MirFlickr dataset. Performance metrics Accuracy and Normalized Mutual Information (NMI) are shown. Paired t-tests are performed and the symbol † indicates that MMRSC is significant better than the corresponding algorithm at $p-$value< 0.05. The best performance is indicated in bold.

Dataset	MirFlickr	
Method	Accuracy (%)	NMI (%)
ConcatKmeans	28.5±3.2 †	13.3±4.8†
ConcatNMF	31.4±3.7 †	16.4±4.5†
ConcatSC	35.7±2.5 †	22.2±3.3†
ConcatGraphSC	33.4±2.5 †	18.7±2.5 †
CollNMF	31.5±2.0 †	17.1±2.1 †
MultiNMF	24.0±0.9 †	12.0±2.3 †
CONMF-W	21.0±1.3 †	6.6±0.6 †
CONMF-B	36.6±3.6	21.5±3.1 †
MMRSC	**37.9±1.9**	**23.2±1.3**

accuracy of 0.21. As we mentioned before, the drawback of CoNMF is that it is impractical to select the best performing coefficient matrix, thus limits its application. MMRSC significantly outperforms CoNMF-B for the NMI metric, and also has a better performance than CoNMF-B for the Accuracy metric. It shows that on the heterogeneous dataset Mirflickr, MMRSC can achieve a better performance.

4 Preservation Value for Text

The aim of Text Preservation Values (TPV) Assessment is to answer the questions: (1) Which factors of digital textual contents (such as emails, messages, documents, news articles, publications, etc.) drive human decisions in preserving for future use? (2) Which factors of archived / history digital textual contents trigger the most of human reminiscence? (3) How does the impact of these factors change over time?

The first two questions are related, covering two perspectives of presrvation decisions. The third question guides further insight into the problem. For all three questions, the answers greatly depend on domains of applications, cultural and educational background, user preferences, business policies, economic conditions, etc. In this project, we focus on TPV related to persons, organisations, and **entities** in more general. We also limit the study to the texts that are related to some **situations** of the entity, that is, some events happening to the entity, such as personal wedding, an endeavour (a project which the organisation participates in, personal education achievement, etc.), to a social event relevant to the entity (a visit to a concert). Projecting the texts into some situations allow us to study the different requirements and features for TPV in more intuitive and easy-to-reason way. In this section, we focus on one aspect of preservation value assessment: Deciding about preservation with respect to the *profile* of an *entity* of interest. In other words, we aim to assess the preservation values of text related to an entity of interest, by seeing how much it contributes to the summarizing of the entity's profile.

As preservation value for text of different types of entities and situations are very different, also because of characteristics of textual data in each context, it is infeasible to design a generic framework to assess preservation values that work in all cases. In this deliverable, we chose to focus on three different domains of situations, and study the TPV assessmnet accordingly:

1. **Academic**: Academic situation involves activities of a person in academic communities. Some examples include: Giving lectures, going to conferences, collaborating in an academic network (e.g. visiting institutions,...). In this work, we consider the situation of a scholar attending scientific conferences and collecting relevant knowledge.

2. **Business**: The business situtation we study here involves the activities of an enterprise in setting up and running an e-commerce project.

3. **Public figures**: Public figures are entities frequently (or once frequently) appeared in social media such as celebrities, politicians, popular organisations, or even one public topic that itself become a concept such as a revolution, etc.

We conduct studies for each individual domains, which are reported below. We also report one application of the preservation value assessment for public event from social media, reported separately in Section 5.1.

4.1 Academic Domain: Survey for Conference Profile Preservation

Here, we conduct a study on the preservation requirements in academic situations. The aim of this study is to understand, from the scholars' points of view, the need for preserving the individual as well as collective texts in academic activities. We choose one of the typical academic activities - attending scientific conferences - as the primary subject of this study. The scientific conference situation we consider incluCdes a broad range of activities: submitting scientific results for the publication, scheduling the travelling, social networking, and taking scientific notes. Such activities often result in many textual data generated, or curated: Papers, draft submissions, slides, travelling notes, program schedules, workshop materials, etc. It is interesting to observe, how scholars perceive the preservation values of such textual data before, during and after the conference time.

We design a survey that targets subjects in academic world, including professors, researchers, PhD and Master students, assistants, and people who are occasionally engaged in activities related to scientific conferences. The survey was written in English and disseminated through networks of academics of the ForgetIT partners, also through the participation of the partners in some scientific conferences: the 2015 Conference on Empirical Methods on Natural Language Processing (EMNLP 2015), and the ACM Conference on Information and Knowledge Management (CIKM 2015). Privacy is respected and not of a concern here, as it is not necessary to know personal information of the subject, except classification questions such as profession (e.g. professor or PhD student), age groups. The scale is adapted according to a previous survey conducted in Work Package 2 ([Logie et al., 2014], Section 5).

4.1.1 Method

The survey was conducted online using the Survey Monkey[6] platform. Dissemination was done via mass emails to chosen groups of academic networks of the partners. In some cases, when partners attended a scientific conference, the survey was printed and handed out within the social network contacts during the conferences, in order to increase the number of valid responses (as the impression about the conference is still fresh).

The survey has two main parts:

- *Reminiscence:* The subject is asked to recall aspects of the conference situation, such as programs, social networking activities, sessions, travelling activities. The aim is to understand what dimensions are important for a conference profile, from the scholar's points of view. It also enables the user recalling of the events, so that they can proceed with the second part (preference) of the survey more easily.

- *Preservation Preference:* The subject is asked for their preference in re-organizing texts after the conference time and which information they wish to or not to preserve.

[6] https://de.surveymonkey.com/r/SF6HDJJ

The survey was designed to take 5-15 minutes to complete (trial attempts from the survey designers took 5 minutes), depending on experiences of the subject (for instance, scholars who attended more conferences will find it longer to recall a specific one). For the reminiscence part, the subjects are asked to recall a detail (e.g. the banquet), and to respond between "instant recall" and "cannot recall". The subject is guided to budget approximately 20 seconds to claim an answer as "recall with some efforts". For the preference part, the subject is asked to imagine a compacted profile of the situation (e.g. to make a backup data related to the event from their computer), and need to choose which data to be included. In the classification questions, besides demography and profession questions, one question is asked to classify the subject's experience on the situation ("how many conferences have you visited ?"). In total, there are 10 questions.

4.1.2 Preliminary Results

Options	Number	%
Professions		
Permanent (Full professor, tenure-track)	2	6.25%
Researcher (assistant prof., postdoc) / professionals in R&D industry	10	31.25%
Graduate students / PhD candidates	17	53.13%
Research assistant, Master student	2	6.25%
Other	1	3.13%
Age Range		
Less than 25	1	3.23%
25 - 30	10	32.26%
30 - 39	13	41.94%
40 - 50	4	12.90%
Above 50	3	9.68%
No. of Conferences Attended		
1 - 3	6	18.75%
3 - 5	11	34.38%
5 - 10	7	21.88%
More than 10	8	25%
The First Conferences Attended		
More than 10 years ago	7	22.58%%
5 to 10 years ago	11	35.48%
1 to 5 years ago	9	29.03%
Less than 1 year ago	4	12.90%

Table 7: Number and Distribution of respondents' profile and experiences

Respondent Distribution. There were 32 valid responses at the time of writing. Most of the responses were in September and October 2015, and were graduate students and researchers (53.13% and 31.25% respectively,see Table7). The majority of the respondents are in a relative early phase of their careers, judged by the age range (74% are in the age from 25 to 39) and their experiences in scientific conferences (75% attended less than 10 conferences, 64.51% attended the first conference 5 to 10 years ago). This bias is probably due to the way the survey was disseminated as well as to the contacts

in academic communities of the partners. We acknowledge that this bias influences the result of preservation preferences for conference situation.

Conference Reminiscence. In this part, the subjects are asked to try recalling the details of the conferences they attended. We choose two special conferences: The first and the most recent conference the subjects attended. This enables us to contrast the effects of emotion and the retention onto the memory, i.e., detailed of the first conference are hypothesized to be retained better due to its emotional impression, while detailed of the last conference are retained because of the relative freshness of the memory. The details can be roughly classified into three groups:

1. Conference Program: This includes details such as time of the conference, venue location, schedules, sessions, best paper awards announcement, conference workshops, etc.

2. Social Details: This includes recalling social network contacts, social events such as banquets

3. Personal Aspects: Examples are personal presentation, notes during sessions, travelling photos, accommodations, etc.

Figure 1 shows the result of to which degree users recall different details of the first conference (left) and of the most recent one (right). Here we map options (from “cannot recall” to “instant recall”) to numerical values on the scale 0-2, and used weighted average rating as a unified score. It can be seen that the distribution of recallabilities are quite consistent among different aspects, regardless the conference visited are the first or the last one (although every detail of the last conference are easier to recall, which is expected). Among the most memorable attributes of the conference are the venue, accommodation information and personal experiences of presentation in the conference. Social aspects such as contact information (e.g. your research acquaintances met at the conference), or banquet events also exhibit a high retention in human’s memory about the past conferences. This is astonishing if compared to the low recall scores of the primary information such as conference schedules, sessions, etc. However, we believe that this also reflects the main motivation of the conference attendance, especially for scholars at their early career phase (see above for the biased distribution in the subjects’ profiles and background), as they often seek broadening or maintaining their visibilities in the communities. This finding is beneficial to the design of the preservation for conference situation-related data, distinguished from other types of datasets.

Preservation Preferences. Continuing from previous analysis, in Table 8 we show the summary of responses for questions regarding directly to the motivation of keeping conference-related data. The majority (77.42%) collect and store material of the during the conference time, and also many (70.96%) get back to such materials frequently or occasionally for work purposes. By computing correlation between this movation and the subject’s background, we found out that for scholars who attend more than 10 conferences, 67% collected conference material during the sessions, and only 25% decides to get back to the material. In contrast, for scholars who attend less than 10 conferences,

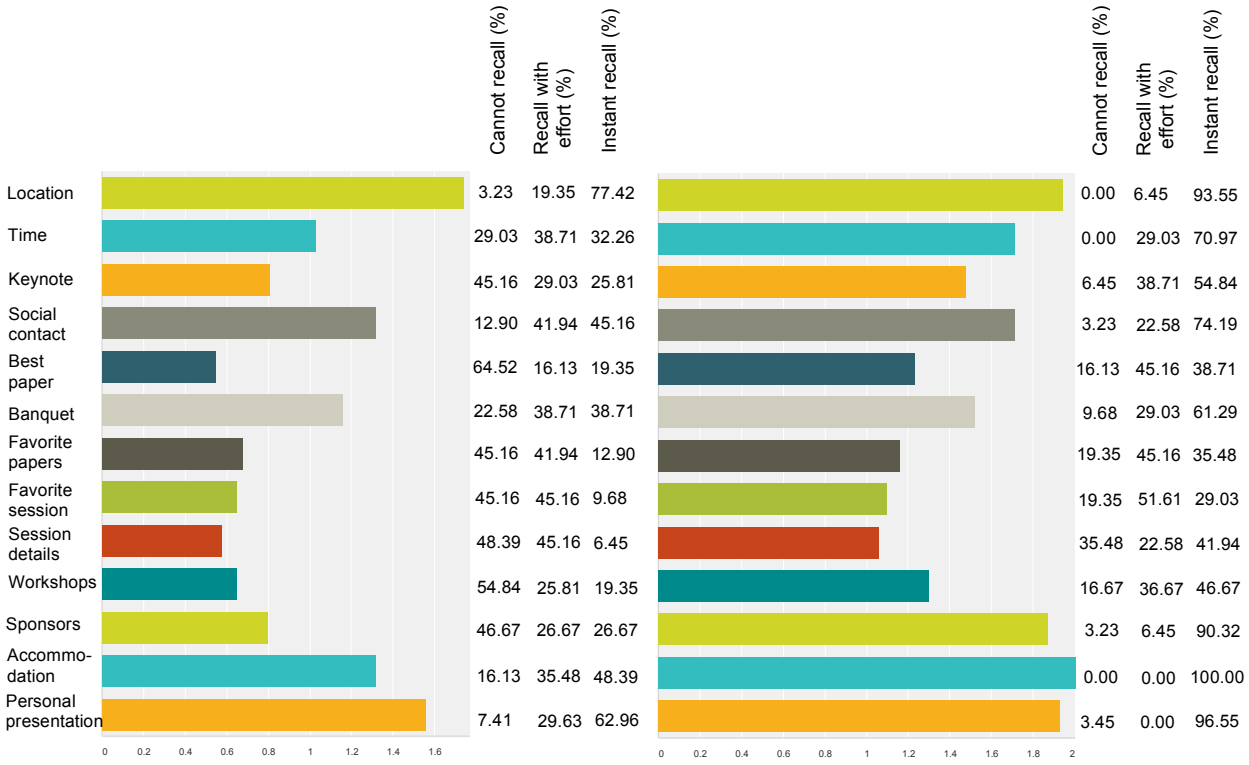

Figure 1: How people recall different details of their first conference (left) and their most recent conference (right)

74% collected and organised material both during and after the sessions, and 55% decide to look back to their collection afterwards. One possible explanation is that the motivation towards preservation of conference-related data is affected by the career needs of the scholar. Further studies would be required to verify this hypothesis.

Questions / Answer	Number	%
How do you collect and store conference material (papers, photos, slides,..) ?		
Collected during conference time, stored offline in computers / bookmark in browsers	24	77.42%
Collected outside the conference time (downloading papers, photos; rechecking slides, etc.)	19	61.29%
(Why) did you look back at your digital materials of the past conferences ?		
Often, for work purpose	6	19.35%
Occasionally, for work purpose	16	51.61%
Occasionally, for personal purpose (reminiscence, etc.)	8	25.81%
1-2 times only, for mixed reasons	7	22.58%
Never	1	3.23%

Table 8: Responses on Preservation Motivation of Academic Data

Next we study in details the attributes of a desired conference situation preservation. We asked the subject to imagine the context in which they need to compile and consolidate the data to provide a conference profile (for instance, due to the limited capacities of the computer hard disk memory). Each subject is asked to select the information they wish to be able to recover from the consolidated set. The results are shown in Figure 2.

For each of the attributes, the percentages of the three options (from “do not want to include” to “want to include”) are computed. From the Figure, we can see that basic, adminstrative attributes such as conference name, time, location, schedules are of high demand for the confrence profile. More personalized and temporal information such as private presentation drafts, personal schedules is often not needed after the conference, although some information (such as video talks) receive contrasting preferences. Other personalised, work-related documents such as research notes, keynote / tutorial materials are demanded to be included into the conference profile, given that they are re-organised and cleansed.

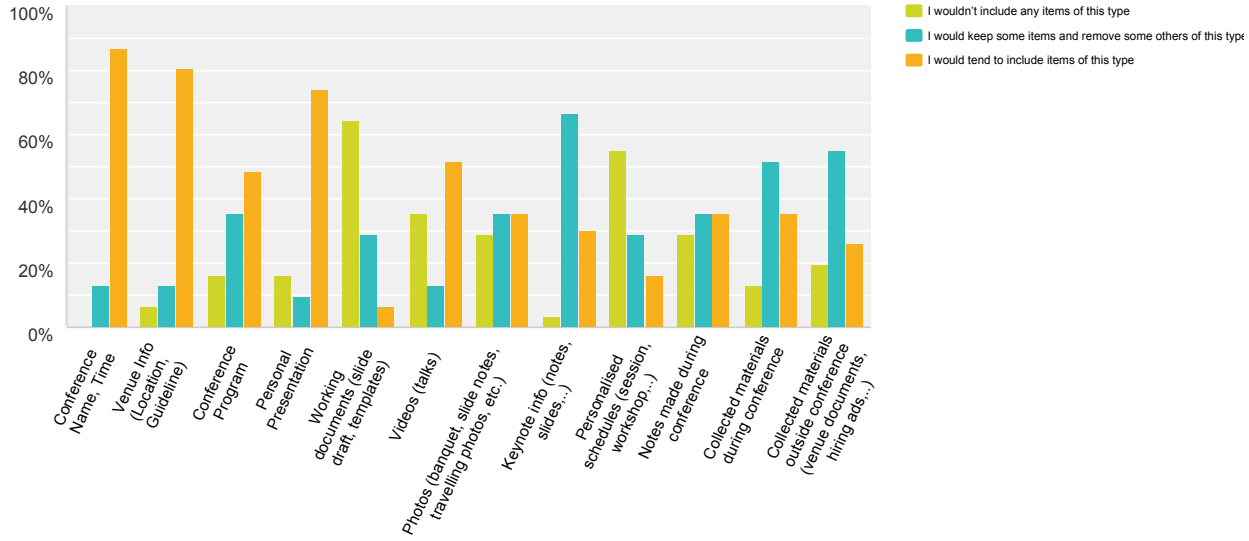

Figure 2: Responses on Preference of Conference Profile

To conclude, the conference situation is a popular activity type in academic, and through our survey, we see some demands for preserving some information of the conference-related data. Interesting findings are the difference in recall of social and professional aspects of the past conferences, and the potential correlation between motivation on preservation of conference-related data and the career needs of scholars. We understand that the survey is still of small size and the subjects targeted are specific, thus the results are hard to generalize. A future direction could be to disseminate the survey into broader communities, and also to incorporate more questions to get better insights into the preservation preferences on this type of data.

4.2 Business Domain: The Fish-Shop Project

In the business domain, we study the preservation scenario for textual assets of an enterprise. The situation we consider involves the setup and maintenance of e-commerce Web sites. We use the setting and data conducted in the evaluation of the Work package 10 for the simulation of the Fish-shop project. The details of data setup and user activities are described in the deliverables D10.4 and D2.4. In this deliverable, we only discuss

the data mining perspective. Specifically, we discuss the workflow of modelling data and learning to assess the preservation values of CMIS objects as produced and annotated by users in the Fish-shop project.

4.2.1 Data Model

Fake's Famous Fishshop include a set of e-commerce websites [7] powered by TYPO3 technology. The core contents are product information (fish), as well as news about different types of fishes. As a typical CMS framework, these contents are produced and managed via a dashboard interface (see D10.4). They also follow the schema that allows them to be rendered properly in the website. This schema defines the data model, part of which is extracted and studied in this study. We list here the relevant textual objects from the FishShop data schema:

1. **Page**: A page is a document describes one complete content: A product details, a list of products, news article, or a static text such as the introduction page "About". Each page has one URL in FishShop domain. In TYPO3 schema, each page is stored as a record in the table *Page*. A page can also be set by the owner to be hidden from the web site.

2. **Content Element**: This is the constituent section in a page, such as the body text, header, an address part, etc. In TYPO3 schema, content elements are records stored in the table *TT_Content*.

3. **File**: A file object refers to a physical MIME-typed document embeded in one section (content element) of a TYPO3 page, for example, an Image or a pdf. In TYPO3, files are stored in the table *Sys_File*. The presence of a file in a page is kept in the table *Sys_File_Reference*. The table also stores information to indicate whether the file is used as an anchor to other pages - in this case, it forms an file-sharing links between two pages (see below).

Relationship. Between different objects of the above three types can exist different relations. Here we only list the relevant relations between two Page items, as they are used in our learning workflow (Section 4.2.2): (1) **Parent**: One page (category or parent page) contains other pages in its structure. For instance, in FishShop website, the page "Freshwater Fish"[8] contains the page "Angel Fish"[9]; (2) **Shortcut**: A page redirects to other page; (3) **Link**: A page that contains a link to other page. This relation differs from the Parent relation in that the links appear in the text of the body content, rather than in the structure or meta-data part; (4) **File-sharing**: Some pages contain image or other MIME-type that can also be used in other pages. For example, a product detail uses an image of a fish, which is also be used in other News page about the same fish. This "file-sharing" implicitly indicates the content similarity between two pages.

[7] One example: `http://web2.fish-shop.net/`

[8] `http://web2.fish-shop.net/fish/freshwater-fish/`

[9] `http://web2.fish-shop.net/fish/freshwater-fish/angelfish/`

Event Log. FishShop extends the TYPO3 backend framework to monitor the user activities on all of its contents. Whenever the user uses the dashboard to create, modify, manage, or publish an object, the action is logged and stored in an InfluxDB database[10]. Each record contains the timestamp, action type (annotate, delete, create, etc.), modified text snippet, and the CMIS id of the object. This log facilitates the user evaluation for the organisation preservation scenario that has been reported also in the deliverable D10.4. In this deliverable, we make use of this information to extract some features for each page (see Section 4.2.2 below).

Statistics. In total, we have 87 Page object, among which 16 are News items. They are constituted by a total of 305 Content Elemen objects and 495 MIME files (481 images, 1 video, 4 pdf documents, 2 embeded HTML pages, among others). The log contains 6276 actions aggregated from all study of 10 participants (reported in D10.4), and exclude all machine-specific actions (i.e., only actions with a non-empty user are kept). The graph constructed from different Page relations consist of 158 edges for the 87 nodes (each corresponds to a Page object), among which: 85 of Parent type, 6 of Shortcut type, 32 File-sharing and 35 Link types.

4.2.2 Learning Process

In this study, we aim to assess the preservation of a Page object as a whole, rather than preserving each parts of Content Element or File objects. Following the general categorisation of preservation value(PV) assessment (deliverable D3.3, Section 3.1.4, Figure 4), we devise 5 labels for for the PVs, encoded from 1 to 5. We employ a supervised machine learning approach, where the preservation values are first manually labelled for some Page by some common assumptions, then applied for the others. We employ the Random Forests [Breiman, 2001] model, as it can learn the association rules about the attributes. In organisation setting, this is an advantage, as such rules can offer the first guidelines for designing the policy of the preservation strategies.

To label the PVs of the Page objects, we adhere to the FishShop project situation: It simulates a scenario in which an enterprise aims to set up and maintain an e-commerce website, and its employees and collaborators provide the contents gradually. Throughoutthese activities, we observe and try to predict, the preservation values of the objects, based on their contents as well as on the log of activities the user perform on them. To be abble to provide the training labels for the Random Forests model, we sample some objects and label according to the following assumptions:

- Pages that are generated automatically as part of the framework have lowest preservation values. For instance, the page “Feature” that lists only general features of a TYPO3 website is labelled 1 (*ash* as per Figure 4, Deliverable D3.3).
- Pages only created for testing purpose (e.g., news pages no real contents) are labelled 2 (*wood*).

[10] `https://influxdata.com/`

- Pages that serve as a category, or shortcut to other actual pages are labelled 3 (*bronze*).
- Pages that contain real contents, but are set hidden (e.g., because the contents are obsolete) are labelled 4 (*silver*).
- Pages that contain real contents and appear in the website (e.g., product detail page, news page) are labelled 5 (*gold*).

Features	*Description*	*Features*	*Description*
FishLikes (U)	No. of likes for a Page	Actions (U)	How many actions performed
ContentElem (M)	No. of content elements	Type (M)	TYPO3 type of the page
Hidden (M)	Whether the page is hidden	Time (M)	Creation / last modified time
SubTitle (M)	Whether page has subtitles	List (M)	Portion of lists in the content
(Sub/Title)Len (T)	Length of title / substitle / main body text	Links/Ref (M)	Porition of links / MIME files references in the content
CType (M)	portion of different types of content elements	Tf (T)	the tf-values of different words in the content
In/Out-degree	No. of In- and Out-links in the Relation graph	Sorting (M)	layout index of content elements in the page

Table 9: Features used for Learning Rules from FishShop data

Features. We extract different features, categorized in four groups:

1. **Meta-data**: This corresponds directly to the attributes of the Page object, as extracted from the TYPO3 tables *Page* and *TT_Content*. For example, the length of the title, or the visibility of the page (hidden or not) in the website.
2. **User**: These attributes are extracted from the user actvitites, either from the log (e.g., how many actions the user performed for the page), or from the social add-ons of the website (e.g. the number of Likes for a page)
3. **Text**: We extract and concatenate words from the body, title and subtitle of the page and all of its content elements, and extract the tf-values for each. Here we also manually check and remove the special "stop words" in the context of TYPO3 framework, e.g. "typo3", "cms", etc.
4. **Graph**: This group consists features extracted from the graph of page Relations. It aims to measure the popularity of the page (how many other pages refer to it), or the investments on the page (how many links from this page to others).

Table 9 summarizes the feature groups extracted from the FishShop data. In total, we have 72 features, and the characters *M*, *S*, *U*, *G*, *T* indicate the categories of the features as described above.

Learning Rules. We use 5-fold cross validation to experiment the rule learning using Random Forests. Essentially, the random forests model learns to generate different decision trees (10 in our setting) with bounded depth per each (we set maximum depth to

20), such as the averaged agreements are maximized according to some criterion metrics over all the trees. In our experiment, we use Information Gain Ratio [11] as the metric to optimize and prune the trees accordingly. Figure 3 shows some presentative trees.

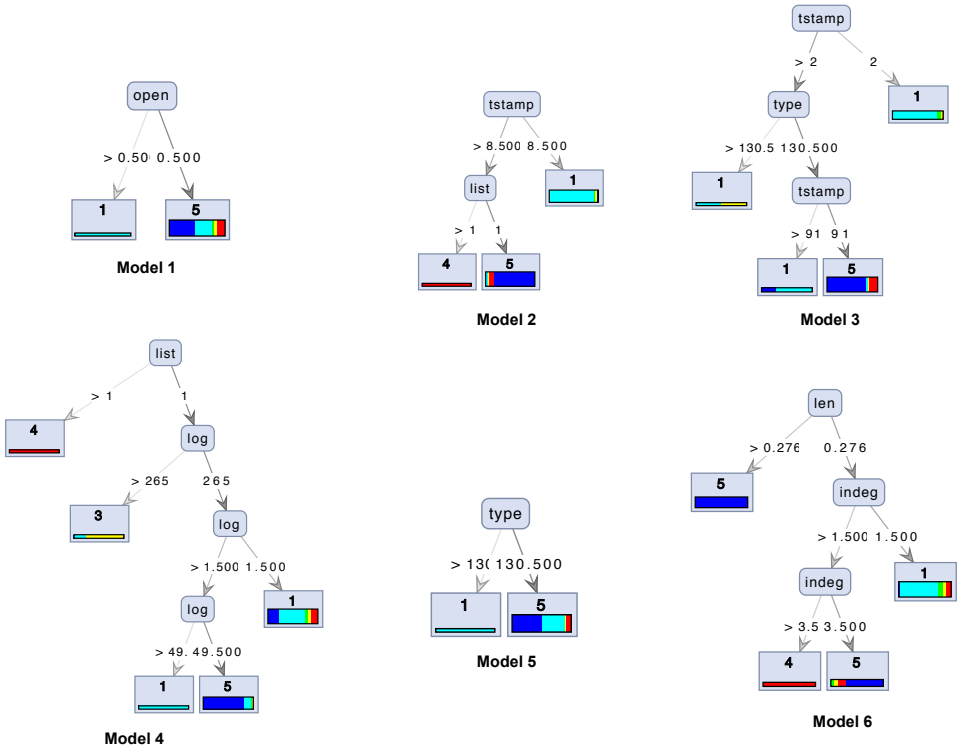

Figure 3: Learned Rules for the Preservation Values of FishShop Pages by a Random Forests Model, implemented in RapidMiner

These trees indicate the learned rules for the preservation strategy. For example, in the tree 6, the corresponding rule is that if the length of the text is greater than 27.6% of the maximum length, then the page should be set as *Gold* for the preservation (5). If not, then we should look into how many other pages refer to the page of interest (in-degree) to make the decision accordingly.

In conclusion, in the business domain, we conducted a study about the preservation for enterprise text assets in situations of setting running an e-commerce project. As the data and the setup is artificially simulated, we aim not to make a conclusion regarding the impacts of the features and dimension of attributes to the preservation strategy. However, this study demonstrates to the concept of learning preservation and rules in the organisation setting, and specially in situations where TYPO3 and CMIS technolgies are involved.

[11] https://en.wikipedia.org/wiki/Information_gain_ratio

4.3 Public Figure Domain: Populating Wikipedia Profiles

In this section, we study the situation of public figures, such as celebrities, politicians, public organisations, even public events. For this type of entities, the situations happening to them (e.g., wedding events, the person's participations in some occasions) are often well-covered in public media such as news streams. To construct a good set of situations for public figures, we need some biography database, which is both of high quality and coverage. In this study, we chose Wikipedia as the database for study the situations of the most popular public figures. In other words, each public figure (entity) corresponds to a Wikipedia page. Note that we do not aim to study all Wikipedia pages, but only ones who refer to a good and popular entities (details in Section 4.3.2).

Wikipedia is contributed by many volunteers worldwide and provides a good summary about an entity. Contents are often presented in chronological order, with important parts in the life of and entity are well organised in sections or lists, and provides references to credited resources such as news articles. In this regard, Wikipedia can be considered as the presevation mirror of the entity's life. By contrasting contents of articles cited by the Wikipedia page with contents of articles mentioning the corresponding entity, we can get insights into what makes a text preservation-worthy. This is the motivation of our study on text preservation values assessment for public figures. In short, given a Wikipedia page about one popular entity, we propose a **news suggestion** to populate the content of the entity, i.e. content of the page. The work has been published as a full paper in ACM conference on Information and Knowledge Managment (CIKM) 2015 in Melbourne, Australia. Below we summarize the highlighted parts of the paper.

4.3.1 News Suggestion Approach

Our news suggestion considers a news article as input, and determines if it is valuable for Wikipedia. Specifically, given an input news article n and a state of Wikipedia,the news suggestion problem identifies the entities mentioned in n whose entity pages can improve upon suggesting n. We propose a two-stage supervised approach for suggesting news articles to entity pages for a given state of Wikipedia. First, we suggest news articles to Wikipedia entities (**article-entity placement - AEP**) relying on a rich set of features which take into account the *gravity* and *popularity* of entities, and the novelty of news articles to entity pages. Second, we determine the exact section in the entity page for the input article (**article-section placement - ASP**) guided by class-based section templates. Figure 4 illustrates these two steps.

Dimensions. We focus on the following four dimensions (discussed in Section 2.1) when assessing news articles with respect to the entity: *Gravity*, *Popularity*, *Coverage & Diversity*, and *Quality*. For the gravity dimension, we make a constraint that it is not sufficient if the news article mentions the entity, but the entity must also be *salient* in the content, i.e., it stays close to the core information of the text. We devise different salience features in our learning model. For the popularity dimension, we measure the global frequency of an entity as compared to others co-mentioned in the article (called the entity *relative*

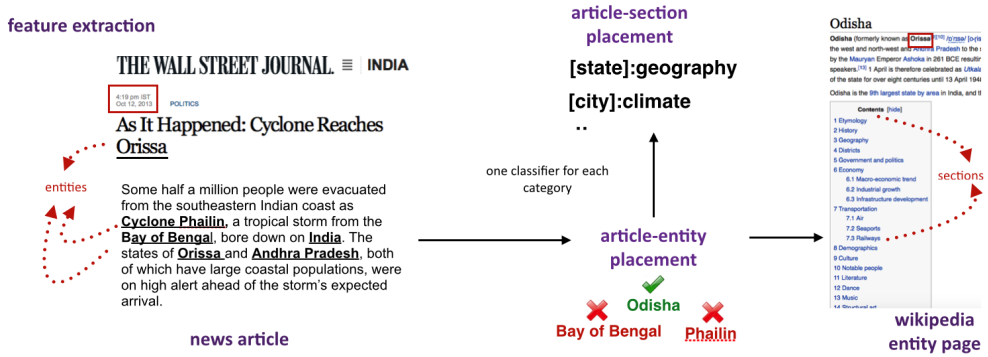

Figure 4: Illustration of News Suggestion Workflow

authority). The intuition is that for an entity that has overall lower authority than its co-occurring entities, a news article is more easily of importance. As for the coverage & diversity dimension, we aim to minimize the redundancy of the news articles compared to the contents presented in the entity profile, or to other existing news articles already referred from the Wikipedia page. In our approach, we devise *novelty* features to address this dimension. Finally, we consider a news article of a high quality, if it facilitates a new fact about the entity in fine-grained level than just some general remarks. In the context of Wikipedia population, we devise *placement* features to determine whether one news should go to one specific section of the page, or it is only relevant to the whole page.

4.3.2 Experiments and Highlighted Results

Datasets and Preprocessing

The datasets we use for our experimental evaluation are directly extracted from Wikipedia entity pages and their revision history. The generated data are publicly available[12]. It contains $73,734$ entities with 27 entity classes selected from DBpedia ontology. For the news articles, we extract all news references from the collected Wikipedia entity pages. The extracted news references are associated with the sections in which they appear. In total there were $411,673$ news references, and after crawling we end up with $351,982$ successfully crawled news articles. We consider year as the interval unit for time t.

Article-Entity Ground-truth. Based on the news reference, news-entity pairs are relevant if the news article is referenced in the entity page. Non-relevant pairs (i.e. negative training examples) consist of news articles that contain an entity but are not referenced in that entity's page. If a news article n is referred from e at year t, the features are computed taking into account the entity profiles at year $S_e(t-1)$.

Article-Section Ground-truth. Also based on the news references, the dataset consists of the triple $\langle(n, e, s)$, where $s \in \hat{S}_c$. Similar to the article-entity ground truth, here too the features compute the similarity between n, $S_e(t-1)$ and $\hat{S}_c(t-1)$.

[12] http://l3s.de/~fetahu/cikm2015/data/

Results of AEP step

For the AEP step, we consider the following baselines: (1) [Dunietz and Gillick, 2014] (denoted B1) only considers salience features; (2) assigns the value relevant to a pair $\langle(n, e)\rangle$, if and only if e appears in the title of n. Figure 5 shows the results for the years 2009 and 2013, where we optimized the learning objective with instances from year t and evaluate on the years $t_i > t$. The results show the precision-recall curve, with the red curve for B1 and the blue curve for our approach (denoted F_e). It is evident from Figure 5 that for the years 2009 and 2013, F_e significantly outperforms the baseline B1. We measure the significance through the t-test statistic and get a p-value of $2.2e - 16$. The improvement we achieve over B1 in absolute numbers, $\Delta P = +0.5$ in terms of precision for the years between 2009 and 2014, and a similar improvement in terms of $F1$ score. The improvement for recall is $\Delta R = +0.4$. The relative improvement over B1 for P and $F1$ is almost 1.8 times better, while for recall we are 3.5 times better.

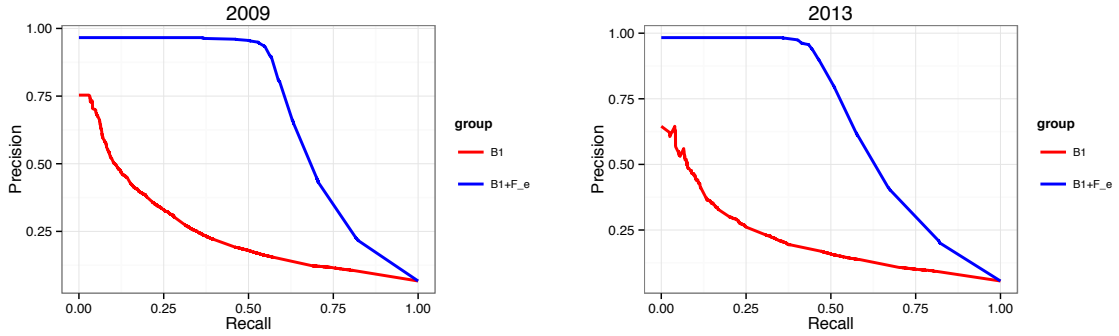

Figure 5: The Precision-Recall Curves in the AEP step

Results of ASP step

For the ASP step, we consider the following baselines: (1) S1, which picks the section from template $\hat{S}_c$ with the highest lexical similarity to n: $S1 = argmax_{s \in \hat{S}(t-1)} \langle n, e, s \rangle$; (2) and S2, which places the news into the most frequent section in $\hat{S}_c$. Figure 6 shows the overall performance and a comparison of our approach (F_s) when optimized using SVM against the best performing baseline S2. With the increase in the number of training instances for the ASP task the performance is a monotonically non-decreasing function. For the year 2009, we optimize the learning objective of Fs with around 8% of the total instances, and evaluate on the rest. The performance on average is around $P = 0.66$ across all classes. Even though for many classes the performance is already stable (as we will see in the next section), for some classes we improve further. If we take into account the years between 2010 and 2012, we have an increase of $\Delta P = 0.17$, with around 70% of instances used for training and the remainder for evaluation. For the remaining years the total improvement is $\Delta P = 0.18$ in contrast to the performance at year 2009. On the other hand, the baseline S1 has an average precision of $P = 0.12$. The performance across the years varies slightly, with the year 2011 having the highest average precision of $P = 0.13$. Always picking the most frequent section as in S2, as

shown in Figure 6, results in an average precision of $P = 0.17$, with a uniform distribution across the years.

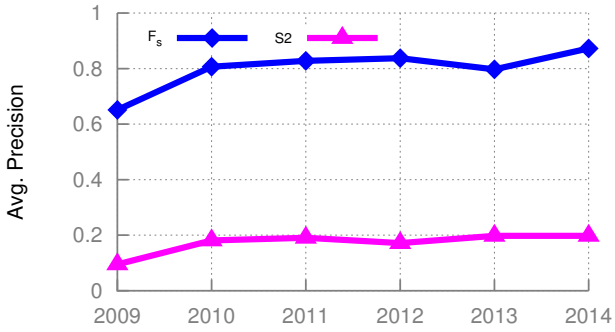

Figure 6: ASP performance averaged for all entity classes

Entity Profile Expansion. The last analysis is the impact we have on expanding entity profiles $S_e(t)$ with new sections. Figure 7 shows the ratio of sections for which we correctly suggest an article n to the right section in the section template $\hat{S}_c(t)$. The ratio here corresponds to sections that are not present in the entity profile at year $t-1$, that is $s \notin S_e(t-1)$. However, given the generated templates $\hat{S}_c(t-1)$, we can expand the entity profile $S_e(t-1)$ with a new section at time t. In details, in the absence of a section at time t, our model trains well on similar sections from the section template $\hat{S}_c(t-1)$, hence we can predict accurately the section and in this case suggest its addition to the entity profile. With time, it is obvious that the expansion rate decreases at later years as the entity profiles become more "complete". This is particularly interesting for expanding the entity profiles of long-tail entities as well as updating entities with real-world emerging events that are added constantly. In many cases such missing sections are present at one of the entities of the respective entity class c.

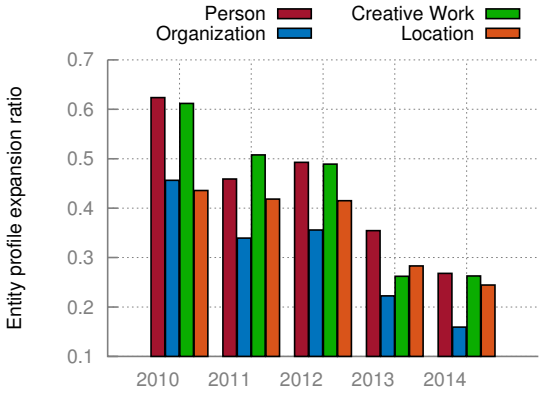

Figure 7: ASP impact on Entity Profile Expansion per different entity classes

Through our ASP approach F_s , we are able to expand both long-tail and trunk entities. We distinguish between the two types of entities by simply measuring their section text length. The real distribution in the ground truth is 27% and 73% are long-tail and trunk entities, respectively.

5 Preservation Value for Social Media

5.1 Collective Memory in News Event Timeline Summarization

5.1.1 Introduction

In this work, we continue our study in preservation values of public text such as news, inspired by the collective memory study (reported in Section 4, [Kanhabua et al., 2014]). In assessing preservation values of news, one interesting type of situations are past events, especially events onced high-impact such as natural disaster, or mass crime incidents such as the Boston Marathon bombing 2013. In retrospect, it is insightful to see what a user remembers and what she might want to re-check about such past events. From a cognitive perspective, for event revisiting, we rather create "memory cues" for real-life remembering [van den Hoven and Egge, 2014]. **Entities** such as persons and locations have been identified as very effective external memory cues [Berntsen, 2009]. Following this idea, we conduct a study that uses entities as a pivot to evaluate the past news events' preservation value. To project the evaluation into a justifiable experimentation framework, we propose to build a prototype of *timeline summarization* for news events, and see how it affects the user digestion and revisiting of the news. We propose to build *entity timelines* with entities as main units of summarization, as depicted in the case of the 2015 Germanwings plan crash (Figure 8). Such summaries can be easily digested and used both as starting points for personalized exploration of event, and for retrospective revisiting.

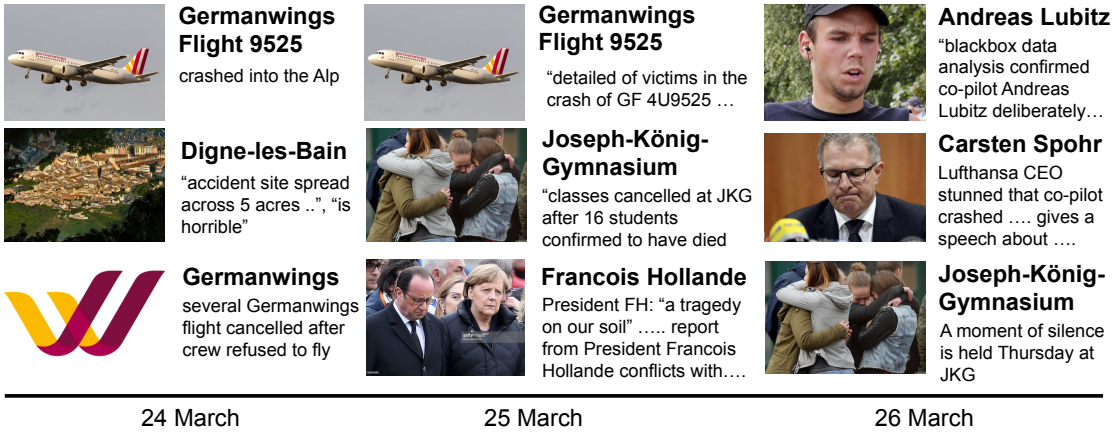

Figure 8: Illustration on Entity Timeline for the 2015 Germanwings Plane Crash Event

For creating an entity timeline,the entities to be used in the summary have to be chosen carefully. In this study, we propose to dynamically combine **salience** with the **informativeness** of entities at a considered point in time. Entity salience, on the one hand, considers the property of being in the focus of attention in a document has been studied in previous work [Boguraev and Kennedy, 1997, Gamon et al., 2013, Dunietz and Gillick, 2014]. Informativeness, on the other hand, assesses the level of new information associated with an entity in a text and can be computationally measured using features derived from

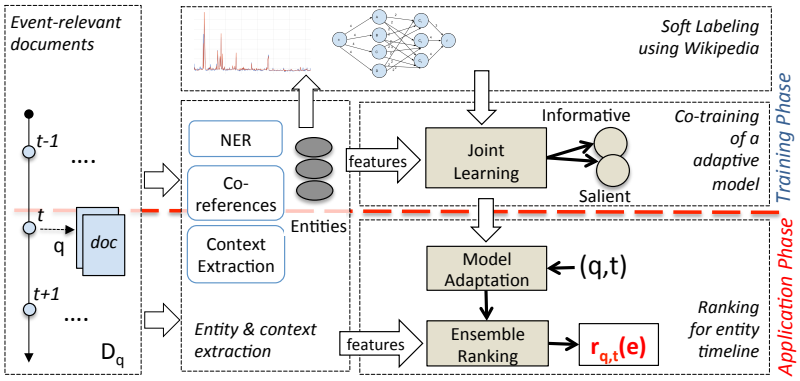

Figure 9: Overview of the Entity-centric Summarization Framework

statistical and linguistic information [Wu and Giles, 2013].

In short, we aim to optimize a trade-off between the in-document salience of entities and the informativeness of entities across documents describing the unfolding of an event. The work has been published as a full paper in ACM Conference on Information and Knowledge Managment (CIKM) 2015. Below are the highlighted parts of the paper.

5.1.2 Entity for Event Timeline: Framework

Figure 9 gives an overview of our entity ranking framework. Given one event q, its reporting timeline, and the set of documents D_q (in practice, D_q can be given a priori, or can be retrieved using different retrieval models) we identify the entity set E_q using our entity extraction, which consists of named entity recognition, co-reference and context extraction. When the event is used for training (training phase), we link a subset of E_q to Wikipedia concepts, which comprises the popular and emerging entities of the event. To facilitate the learning process, these entities are softly labeled using view statistics from Wikipedia, serving as training instances. Although we use popular entities for training, we design the features such that it can be generalized to arbitrary entities, independent from Wikipedia. The next component in our framework is the adaptive learning that jointly learns the salience and informativeness models, taking into account the diverse nature of events and their evolution. Finally, in the application phase, entity and feature extraction are used and against the joint models to produce the final ranking scores. More details are in [Tran et al., 2015b].

5.1.3 Experiments and Evaluations

Datasets. For our experiments, we work with a real-world, large-scale news dataset. Specifically, we use KBA 2014 Filtered Stream Corpus (SC14) dataset[13]. We extract news and mainstream articles from the dataset, consisting of 7,592,062 documents. The

[13] http://s3.amazonaws.com/aws-publicdatasets/trec/kba/index.html

dataset covers 15 long-running news events from December 2012 to April 2013. All events are high-impact, discussed largely in news media, and have their own Wikipedia pages. Each event has a predefined time span (ranging from 4 to 18 days), and is represented by one textual phrase that is used as the initial event query.

Training Data. From 153 (*event*,*day*) pairs, we randomly choose 4 events belonging to 4 different categories mentioned above as a training data, resulting in 39 pairs. To build training entities (i.e. to identify subset of entities to Wikipedia concepts), we apply two named entity disambiguation softwares, WikipediaMiner and Tagme. These are the supervised machine learning tools to identify named entities from natural language texts and link them to Wikipedia. We train the models of both the tools from a Wikipedia dump downloaded in 2014 July, so as to cover all possible entities in the SC14 corpus. We only use entities co-detected by both the tools, resulting in 402 distinct entities and 665 training tuples (*entity*,*event*,*day*). We use the Wikipedia page view dataset, which is publicly available, to build the soft labels for these entities.

Baselines. We compare our approach with the following competitive baselines.

TAER: Dermatini et al. [Demartini et al., 2010] proposed a learning framework to retrieve the most salient entities from the news, taking into consideration information from documents previously published. This approach can be considered as "salience-pro".

IUS [McCreadie et al., 2014]: This work represents the "informativeness-pro" approach, it attempts to build update summaries for events by incrementally selecting sentences, maximizing the gain and coverage with respect to summaries on previous days.

In addition, we evaluate three other variants of our approach (Details in [Tran et al., 2015b]). The first two variants involve only salience and informativeness features for learning. We denote these as *SAL* and *INF*. The third variant linearly combines all salience and informativeness features, denoted as *No-Adapt*. All are trained using RankSVM.

Evaluation Metrics. We consider the traditional information retrieval performance metrics: precisions, NDCG and MAP. Besides, we also use serendipity [Bordino et al., 2013] to measure the informativeness and salience by contrasting the results of one day to previous day of the same event:

$$SRDP = \frac{\sum_{e \in UNEXP} rel(e)}{|UNEXP|} \tag{5.7}$$

where $UNEXP$ is the set of entities not appearing on the previous day, and rel is the human relevance judgment of the entity. This measures aims evaluate the performance of ranking in timeline summarization context, where effective systems do not just introduce relevant, but also novel and interesting results compared to the past [Ge et al., 2010, Bordino et al., 2013]. The relevance part ensures the salience of the entity, while the $UNEXP$ part ensures the informativeness of the entity.

Assessment Setup. We exclude the 39 training pairs from the overall 153 (*event*,*day*) pairs to obtain 114 pairs for testing. For each of these pairs, we pooled the top-10 entities returned by all methods. In total, this results in 3,336 tuples (*entity*, *event*, *day*) to be

assessed. To evaluate the quality of the systems, we employ an expert-based evaluation as follows. 5 volunteers who are IT experts and work on temporal and event analysis were asked to assess on one or several events of their interest. For each event, the assessors were encouraged to check the corresponding Wikipedia page beforehand to gain sufficient knowledge. Then, for each tuple, we add one more contextualizing sentence, extracted from the previous date of the event. If there is no such sentence, a "NIL" string will be presented. We asked the assessors to check the tuple and the two sentences, and optionally, to use search engines to look for more event information on the questioned date. Then, the assessors were asked to assess the importance of the entity with respect to the event and date, in four following scales. **1**: Entity is obviously not relevant to the event; **2**: Entity is relevant to the event, but it has no new information compared to the previous day; **3**: Entity is relevant to the event and linked to new information, but it does not play a salient role in the sentence; **4**: Entity is relevant to the event, has new information, and is salient in the presented sentence. The inter-assessor agreement score for this task is $\kappa = 0.4$ under the Cohen's Kappa score.

5.1.4 Results and Discussion

Method	P@1	P@3	P@10	MAP	SRDP@1	SRDP@3	SRDP@10
Ranking performance from expert assessment							
TAER	0.436	0.315	0.182	0.109	0.315	0.210	0.121
IUS	0.395	0.325	0.236	0.141	0.335	0.217	0.176
SAL	0.493▲	0.423	0.338▲	0.217▲	0.421	0.320	0.240▲
INF	0.480▲	0.436	0.354▲	0.227▲	0.441▲	0.340	0.256▲
MAX(S,I)	0.493	0.436	0.354	0.227	0.441	0.340	0.256
No-Adapt	0.503	0.461	0.320	0.225	0.396	0.338	0.215
AdaptER	**0.546**	**0.485**	**0.368**	**0.264**	**0.507▲**	**0.440▲**	**0.275**

Table 10: Entity-ranking performance using different assessment settings. Symbol ▲ indicates cases with confirmed significant increase, tested against line 1, TAER (first group), and line 5, MAX(S,I) (second group)

Table 10 summarizes the main results of our experiments from the expert evaluation. The results show the performance of the two baselines (*TAER* and *IUS*) and of the consideration of Salience and Informativeness features in isolation with respect to precision. In general, all performances are low, indicating the relatively high complexity of this new task. In addition, as can be seen from this part of the table, even the approach relying on our salience features or informativeness features in isolation already outperforms the two baselines. This is due to the fact that our approach does not consider documents in isolation as the baselines do. Rather, we take a more comprehensive view considering event level instead of document level features via feature aggregation.

Furthermore, the results also show the performance of the non-adaptive combination of salience and informativeness (*No-Adapt*) as well as our approach (*AdaptER*), which uses an adaptive combination of informativeness and saliency. It becomes clear that

an improvement by combining the salience and the informativeness features over the use of the isolated features can only be achieved by fusing the two features in a more sophisticated way: *No-Adapt* does not perform better than the maximum of *SAL* and *INF* ($MAX(S, I)$), it even performs worse in under some metrics such as P@10. In contrast, *AdaptER* clearly outperforms the maximum of *SAL* and *INF* (we achieve 16% MAP improvement), as well as its non-adaptive version for most metrics.

April 15	*April 16*	*April 17*
Boston Marathon **Mass General Hospital** **Boston.com**	**Boston** **Boston Marathon** **Vatican**	**Boston Marathon** **Boston** **Boston University**
- Two bombs exploded near the finish of the Boston Marathon on Monday, killing two people, injuring 22 others - At least four people are in the emergency room at Mass General Hospital	Deeply grieved by news of the loss of life and grave injuries caused by the act of violence perpetrated last evening in Boston, His Holiness Pope Francis wishes me to assure you of his sympathy . . .	- FBI confirmed that pressure cookers may have been used as explosive devices at the Boston Marathon. - The third victim was identified Wednesday as Boston University graduate student Lingzi Lu.
Boston Marathon **Marathon Bruins** **New York City**	**Pope Francis** **Vatican** **Boston Marathon**	**FBI** **Boston University** **Lingzi Lu**
- Two bombs exploded near the finish of the Boston Marathon on Monday, killing two people, injuring 22 others - The NHL postponed the Boston Bruins' Monday hockey game due to the bombing	The Vatican sent a telegram to Boston Cardinal on Tuesday, in which Pope Francis expresses sympathy for the victims of the marathon bombings. . .	- FBI confirmed that pressure cookers may have been used as explosive devices at the Boston Marathon. - The third victim was identified Wednesday as Boston University graduate student Lingzi Lu.

Table 11: Top-3 entities on Boston Marathon Bombing 2013 learnt by TAER (top) and byAdaptER (bottom)

Besides precision, we also consider serendipity (SRDP) as a complementary measure in our experiments, as discussed above. This metric measures how likely the approach brings unseen and interesting results to the user. Under SRDP, our approach outperforms significantly both the baseline and the maximum of *SAL* and *INF*. We achieve 14% improvement of serendipity at top-1 entities, and 29% at top-3 entities. Thus, our top-retrieved entities do not only cover relevance, but are also more interesting, often unseen on the previous day (contributing to more informative results).

Anecdotic Example. In Table 11, we show one example of top-selected entities for the event "Boston marathon bombing 2013". Additionally, we show some selected sentences covering the entities, to enable the understanding of the entities' roles within the event on the presented days. As can be seen, the timeline corresponding to *TAER* approach (upper part) gives more salience credits to entities frequently mentioned throughout the news (such as Boston marathon), keeping them in high ranks throughout the timeline. The approach is not responsive to less salient but interesting entities (such as Pop Francis, a rather unrelated entity to the event, but get involved via his condolecence and activities to victims of the bombing). On the other hand, using an adaptive ranking with informativeness incorporated, the resulting entities are not just more diverse (including related events such as Marathon Bruins), but also expose more new and emerging information.

To conclude, in this section we reported one application to assess the preservation values

of news articless via the presence of salient and informative entities in its content. Analysis in the context of timeline summarization task has demonstrated the effectiveness of our proposed approach.

5.2 Learning to Rank Memorable Posts in Facebook

In D3.3 we presented a study for collecting ground truth collection, together with the first data analysis and presented a list of features which could be useful for learning to rank memorable posts. We extended our previous study by collection a larger datasets using crowd sourcing to proof that our set of top features can improve ranking on a larger dataset. Further we tried a personalized ranking method using k-nearest neighbour to improve ranking and reducing online processing costs. For the motivation of our work we refer to D3.3.

Dataset. In order to have a sufficient sized dataset for the learning to rank approach, we collected a second dataset using crowd sourcing. The task for the workers was to evaluate their Facebook profiles and rate their Facebook posts using 5 scale where 1 is for not relevant and 5 point is very relevant. To complete the task, the user had to evaluate at least 100 posts and have an Facebook account which is at least 4 years old. Each user got 25 posts randomly selected from each year, from 2014 back to 2010. In cases where the users evaluated more than 100 posts or the Facebook profile of the user had less than 25 post for each year, they got older posts to evaluated. In total we have 466 users from 72 different countries. The task was completed in average by about 102 seconds. In average each user evaluated at least 100 (total 466,000 posts).

Personalized Ranking and Results.

So far we considered a general ranking model learnt for all the users (see D3.3). However, in search domain, a recent study has shown that it is beneficial to build query-dependent ranking models, as queries significantly differ from each other [Geng et al., 2008]. This latter study proposes to use k-Nearest Neighbor (kNN) method so that for a given query first its nearest neighbors are found in the training set and then a customized ranker is learnt using only these neighbor instances. Analogously, in our setup, it is natural to hypothesize that similar users may have similar motivations/instincts while deciding on the memorable posts. Hence, we also apply a kNN based strategy to build more personalized ranking models. We represent each user with a vector of three key features, namely, the number of posts, number of friends, and number of connections among the users friends, which may reflect the coherence in the users network. We anticipate that these features best capture the activity level of a user in a social web application, and users with similar activity patterns can exhibit similar behavior while deciding on the memorable posts. To determine the nearest neighbors of a user, we compute the Euclidean distance between the pairs of these feature vectors, and choose the ones (k of them) that yield the smallest distances. Then, for each test user, only these k nearest neighbours (and their posts) are used to train the RankSVM algorithm.

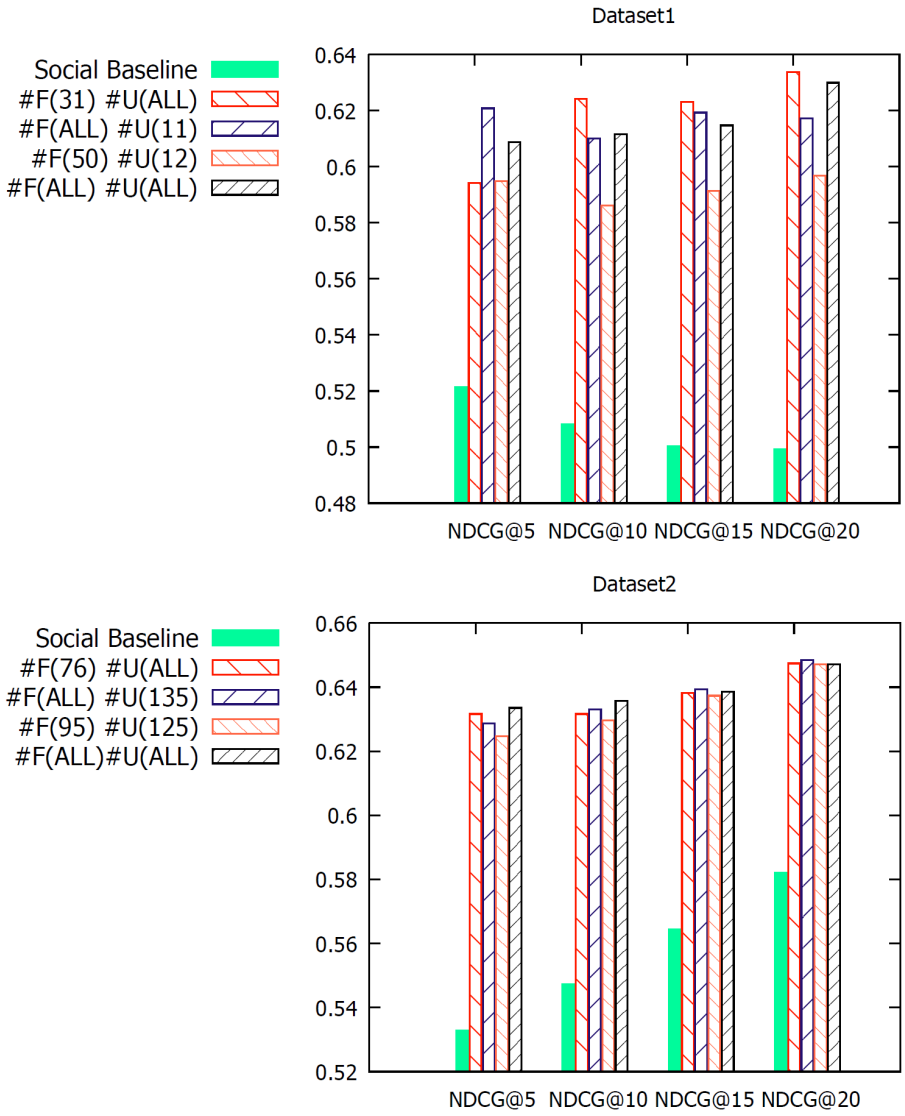

Figure 10: The Effectiveness of general and personalized ranking models.

The Effectiveness of general and personalized ranking models are presented in Fig. 10. The first observation is that all the models are performing better than the social baseline using features as no. of likes, no. of comments, and no. of shares.

The number of features selected by the GAS (see. D3.3) method are conducted with #F and the number of users selected by the kNN for the personalized ranking are conducted with #U. The results of the general ranking using all the users (466) and all the features (135) are represented as ALL. Here we can observe that using only feature selection gives the best results for NDCG@10. NDCG@15, and NDCG@20. In the larger dataset 2 we can observe that our models are very close to the results by using all features and all users. In particular we can observe that using only personalized ranking #F(ALL)#U(135) gives us the best result for NDCG@20. Further we can also observe that by combin-

ing the personalized ranking with features selection can be at least as good as using #F(ALL)#U(ALL). This reduce the cost of learning a model since we use less number of features and instances for training. In web search domain, building a model for each query can imply prohibitive online processing costs, as the users typically expect search result less than a second [Geng et al., 2008]. In our case, this would be less of a concern; as ranking the memorable posts is not an everyday task for a user, but an application that is most likely to be executed periodically, such as de-fragmenting your hard-drive. Hence, the additional processing latency for online model building can be tolerated by the users, for the promise of a better final ranking. Furthermore, it is still possible to improve the efficiency using offline pre-processing techniques, such as clustering, as proposed in an earlier work [Geng et al., 2008]. Finally, by looking to the top-features we could observe a high overlap between them and the features presented in D3.3.

5.3 Analyzing and Predicting Privacy Settings in the Social Web

In the context of WP3, it is important for developed techniques to hold the capability to express required constraints. In this work, we propose to support the users by choosing the right privacy setting by adding new content to their summary. We investigated this problem in our paper [Naini et al., 2015] published at the UMAP conference 2015. In this section, we present a short summary of our findings. For more detailed description we refer to the original paper [Naini et al., 2015]. In this work, we present an approach for supporting users in selecting adequate privacy settings for their posts. This work is based on a thorough analysis on privacy settings on social networks, particularly in Facebook. Targeting a supporting tool that could suggest users preferable privacy settings, we performed experiments for the privacy settings prediction task. By relying on different categories of features that can already be identified at the time of post composition, we were able to achieve a very good prediction performance with a recall and precision of more than 80% on average.

Dataset and Analysis. In this section we present the two datasets which we also used for the learning to rank memorable pots 5.2 and also D3.3. In Table 12, we summarize the characteristics of both datasets.

Experimental Setup:

Target classes. To build a predictor with reasonable accuracy that can be employed in

Table 12: Datasets.

	Dataset 1	Dataset 2
No. of users	45	649
No. of posts	26,528	769,205
Avg. no. of posts per user	602.431	1,185.215
Variance no. of posts per user	545,343	5,484,176
Min no. of posts per user	13	100
Max no. of posts per user	3,176	30,715

Table 13: The list of features used for the privacy prediction task.

Feature	Description	Feature	Description
Post metadata		**Context**	
has(message)	post has a message	sendFromMobile	post sent from an mobile application
length(message)	length of the message	dayTimes	(morning, afternoon, evening, night)
norm(length(message))	length normalized per user	sendAtWeekend	post sent during weekend
has(story)	has a story	**Sentiment**	
length(story)	length of the story	negative	the negativity score of a post
norm(length(story))	length normalized per user	positive	the positivity score of a post
has(description)	has a description	objective	the objectivity score of a post
length(description)	length of the description	**Users**	
norm(length(description))	length normalized per user	no_posts	total number of posts of a user
has(link)	post includes a link	no_friends	total number of friends of a user
has(icon)	post has an icon	gender	gender of the user
has(caption)	post has an caption	age	age of the user
type	type of post	country	country of the user
status_type	status_type of a post	education	the education level of the user
icons	describes user activity	**Keywords**	
tagged users	users tagged in a post	words_family	contains word from the list
Word vector		words_friends	contains word from the list
bag of words	top-1000 words using tf/Idf	words_work	contains word from the list
		words_holiday	contains word from the list
		words_travel	contains word from the list

a practical setting, we opt for building a binary classifier and predicting whether a post has low or high privacy at an abstract level, rather than assigning each post to one of the privacy levels described in Section 3. We assume that posts that have the privacy setting EVERYONE or FRIENDS_OF_FRIENDS are in the class *Low_Privacy*, as they are visible to a very general audience. In contrast, the posts with the setting ALL_FRIENDS, SELF and CUSTOM are said to be in the class *High_Privacy*, as the user has the intention of sharing the post with a specific audience, i.e.; with only her friends, which can be the most typical case in a social platform, or even with a certain subset of them.

Table 14: Classification results using all the features.

Naive Bayes						
TP Rate	FP Rate	Precision	Recall	F-Measure	AUC	Class
0.640	0.255	0.715	0.640	0.675	0.780	LOW_PRIVACY
0.745	0.360	0.674	0.745	0.708	0.780	HIGH_PRIVACY
0.692	0.308	0.694	0.692	0.691	0.780	Avg.
REPTree						
TP Rate	FP Rate	Precision	Recall	F-Measure	AUC	Class
0.810	0.191	0.809	0.810	0.810	0.887	LOW_PRIVACY
0.809	0.190	0.810	0.809	0.809	0.887	HIGH_PRIVACY
0.809	0.191	0.809	0.809	0.809	0.887	Avg.

Features. In our experiments, we use features from six different categories (see Table 13).

Classifiers and evaluation metrics. We apply the well-known classification algorithms NaiveBayes[John and Langley, 1995] as well as a fast decision tree learner, REPTree [Witten and Frank, 2005][Breiman, 1996].

Results and Discussions. In Table 14, we compare the prediction performance for NaiveBayes and RepTree classifiers. The average TPR (i.e., accuracy) of the NaiveBayes predictor is 0.692, which is better than the random baseline with 0.5 accuracy (as we have a balanced dataset). Moreover, when predicting the High Privacy class, the classifier has a higher TPR (i.e., 0.745). This is useful in practice, as predicting a highly private post as public is more dangerous (as these are the cases where the information is exposed to a larger audience than intended) than vice versa. The overall performance of the RepTree classifier is even more impressive, as it yields an accuracy of 0.809 for both classes (and, on the average). For this classifier, average F-measure and AUC metrics are also over 0.80. These findings reveal that it is possible to predict the privacy class of a post with good accuracy, and such a predictor can serve in suggesting the privacy setting of a post when it is first created.

6 Managed Forgetting in Applications

So far in this deliverable and in previous ones of WP3, we have discussed a number of studies and methods to realise the managed forgetting in different scenarios and applications. We have reported some prototypes of summarization photo collections based on preservation values, or personal contents in social media (Section 5, [Kanhabua et al., 2015]). In Section 3, we also discussed the components of photo preservation value assessment as in the context of the "Personal Photo Selection" application.

In this Section, we described two other highlighted applications, which realise the managed forgetting methods: Memory Buoyancy for decluttering semantic information spaces ,and Preservation Value calculation in the Semantic Desktop. The practical issues of some of these applications have been detailed in other deliverables (such as in [Maus et al., 2015a]), and here we only discuss the learning procedure used in such applications.

6.1 Memory Buoyancy in Decluttering Semantic Information Spaces

In this Section, we discuss another application of managed forgetting, which uses Memory Buoyancy to declutter the information spaces. The motivation is based on the fact that finding and re-finding documents in personal and collaborative spaces becomes both more crucial and challenging, due to the growing amount of content stored. Managed forgetting aims to relieve the manual efforts by automatically computing the memory buoyancy of a document with respect to the user attention. Based on this computation, documents with highest values will be recommended to the user. This is the continuation of our research done in memory buoyancy, which was reported as ongoing research in [Kanhabua et al., 2015], Section 6.1. We improved the propagation model ([Kanhabua et al., 2014], Section 2.2.1) significantly by introducing novel machine learning methods for memory buoyancy propagation in heterogenous graphs. We also conducted more experiments to evaluate the contributions of all aspects in our components, qualitatively and quantitatively. Our results have been accepted as a full paper in the 2016 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR' 2016). Below we give the highlights of the paper.

6.1.1 Overview

Following up re-access model ([Kanhabua et al., 2014], Section 2.2), we propose to rank documents via two steps. In the first step, we mine the activity history and devise a memory buoyancy scoring function based on the recency and frequency, so that more recently and frequently accessed documents get higher memory buoyancy scores. This step is based on different time-dcay models ([Kanhabua et al., 2015], Section 2.1). In the second step, a propagation method is used to identify highly connected documents, and propagate the activity-based scores of documents along the connection. A simple propagation method was proposed in [Kanhabua et al., 2013] (Section 2.2.1), however it

failed to identify the different contribution of different types of document relationships on the memory buoyancy propagation. In our new revised system, we make a significant progress by introduce a machine learning framework that can learn the contributions of individual relations and combine them automatically. In this deliverable we focus on this learning method, which is detailed as follows.

6.1.2 Machine Learning Framework of Memory Buoyancy Propagation

Preliminaries. Before we can discuss our propagation learning framework, we need to introduce some formal notations. A semantic information space is a collection of documents or resources and is denoted as D. A document or a resource d can be of different types (photos, office documents, folders, web pages, etc.) and has different attributes (e.g., titles, authors and creation time). Given any two documents d_1 and d_2, there can exist multiple relations with different semantics. For instance, d_1 and d_2 are both created by the same author, d_1 is the containing folder of the file d_2. Relations can be associated with some scores indicating the strength of their relation, for instance, the cosine score for a content similarity. Let R denote a set of all semantic relations. For each pair (d_i, d_j), we have an $|R|$-dimensional vector $X_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ij|R|})^T$, where $x_{ijk} \geq 0$ represents the score of the k-th relation between d_i and d_j, $x_{ijk} = 0$ if the d_i and d_j is not connected by the relation. Usually, the number of relations is small compared to the number of all documents in the information space. The collection of relation scores $X = \{X_{ij}\}$ forms the weights of edges in a multi-graph, where nodes are all documents in D, and each edge corresponds to a semantic relation.

Problem. With these notations, the problem can be formalized as follows. Given a collection of documents D, a set of relation scores X, time of interest t, and an activity history L_t corresponding to a user u, or to a group of users U, identify documents with highest importance with respect to u's or U's task and interest at time t.

Propagation Process. In this new version of propagation model, we treat the process that the user finds the important document (re-access model, [Kanhabua et al., 2013], Section 2.2) as a Markov process, when she recalls and searches for important documents via the related resources. For each pair of connected documents (d_i, d_j), we define the transition probability from document d_i to d_j as:

$$p_{ij}(w) = \begin{cases} \frac{\sum_k w_k x_{ijk}}{\sum_l \sum_k w_k x_{ilk}} & \text{if } X_{ij} \neq \emptyset \text{ and } L_{d_j,t} \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \tag{6.8}$$

where w is the weighting vector for the semantic relations in R. The condition $L_{d_j,t} \neq \emptyset$ ensures that the propagation has no effect on the documents that have not been created before the time t, i.e., no propagation to the future. Similarly, the indices l's run only over the documents d_l with $L_{d_l,t} \neq \emptyset$. Consequently, we have $\sum_j p_{ij} = 1$ for all documents d_i. In practice, to avoid rank sink when performing the propagation, if a document has no relation we assume a dummy edge from it to all other documents with zero probability.

Next, we describe our propagation framework. Let P be the transition matrix of documents

in D, we follow the PageRank model to define the propagation as an iterative process:

$$\mathbf{s}^{(n+1)} = \lambda P^T \mathbf{s}^{(n)} + (1-\lambda)\mathbf{v} \tag{6.9}$$

where $\mathbf{s}^{(n)} = (s(d_1,t), s(d_2,t), ..., s(d_m,t))$ is the vector of documents' memory buoyancy values at iteration n, (m is the number of documents appearing in L_t), $\mathbf{v}$ is the vector of values obtained by an activity-based scoring method, and λ is the damping factor.

Learning Framework. The aim of the learning is to identify the weights $w_1, \ldots, w_{|R|}$ of the semantic relations with which we obtain the best prediction of document rankings. In this work, we propose to exploit the activity history to learn the optimal w. In particular, we simulate the navigation of the user at each time points t' in the past, and compare the computed ranks of the documents with the ranks based on the frequency of access in the time point $t'+1$. The idea is to learn w so as to minimizes the number of mis-ranked pairs (d_1, d_2), i.e. $s(d_1,t') > s(d_2,t')$ but d_1 has been accessed less than d_2 until $t'+1$.

Formally, we define the label $y_{ij} = s(d_i,t') - s(d_j,t')$ and the groundtruth $\hat{y}$, $\hat{y}_{ij} = -1$ if d_i has less access than d_j at $t'+1$ and $\hat{y}_{ij} = 1$ otherwise. We learn w by solving the following optimization problem:

$$\min_w F(w) = \|w\|^2 + \theta \sum_{(d_i,d_j)\in A} h(y_{ij}) \tag{6.10}$$

where A is the training data, θ is the regularization parameter that controls the complexity of the model (i.e., $\|w\|^2$) while minimizes the mis-ranked pairs in A via the loss function h. In this work, we apply the simple hinge loss function: $h(y) = \max(0, 1 - \hat{y}.y)$. The Equation 6.10 then can be solved using the well-known supervised PageRank framework [Backstrom and Leskovec, 2011], and can be done efficiently gradient descent-based learning paradigm. More details can be in our paper at [Tran et al., 2016b].

6.1.3 Experiments

We continue our experiments on the two datasets of L3S wiki backup and PIMO desktop collection ([Kanhabua et al., 2015], Section 6.1.1). As compared with earlier experiments reported in [Kanhabua et al., 2015], in this revision, we evaluate our system against the following baselines:

Recency-Frequency: These baselines use values of the activity-basd scoring functions to provide the final ranking, without propagation. This includes the two recency-based methods MRU and Ebb, and their frequency-based variants, denoted by FMRU and FEbb (details in [Tran et al., 2016b]).

PageRank: This baseline ranks the documents by their authority scores, estimated in a graph of documents relations. The scores of documents are initialized equally. It can be thought of as the propagation method without the activity-based rankings and the semantics of relation. In our case, we adapt the PageRank algorithm by aggregating all relations between two documents into one single relation, with the weighting score obtained by averaging out all the individual relation weights.

SUPRA: Papadakis et al. [Papadakis et al., 2011] proposed combining the activity-based ranking results with a one-step propagation in a layered framework. The relations are constructed simply by identifying documents accessed in the same sessions. In our scenarios, we define the "sessions" to be one unit time step, which is one hour.

Evaluation based on Revisit Prediction

The first experiment aims to evaluate how well the system performs in the revisit prediction task, i.e., predicting the likelihood that a document will be accessed by the user in the subsequent time point. This is the well-established task in research on web recommendation [Connor and Spitzer, 2015], personal file retrieval [Fitchett and Cockburn, 2012], etc. We evaluate the correlation between the predicted rank of a document at a time point t and the real document accesses at the time point $t+1$. Inspired by [Kawase et al., 2011], we employ the following evaluation metrics

1. Success at 1 (**S@1**): It quantifies the fraction of time points t (from all time points of study) at which the first-ranked documents according to a ranking method is truly accessed at $t+1$. This resembles the Precision at 1 (P@1) metric in traditional IR tasks.

2. Success at 10 (**S@10**): It quantifies the fraction of documents truly accessed in the next time point, from all documents ranked at top 10, averaging over all time points of study in the micro-average manner (i.e., per-document average).

3. Average Ranking Position (**ARP**): This metric starts from the subsequent document access backwards. It computes the average ranking position of accessed documents as produced by a ranking method. The lower the value is, the better the performance of the corresponding ranking system.

Results. The average results over the two datasets are summarized in Table 15. Among the ranking methods, PageRank has the worst predictive performance. This is because it ignores the recency and frequency signals of the documents. Other interesting observation is that for activity-based ranking methods, adding frequency into the ranking function did not really help in revisit prediction: FMRU performs worse than MRU and FEbb performs worse than Ebb in all metrics, although the differences are not significant. At the first look, this contradicts somewhat to previous findings on the influence of frequency in document ranking [Peetz and De Rijke, 2013]. However, analysing deeper, we believe that the cause stems from the fact that a revisiting action typically involves very recent documents, as also argued in [Kawase et al., 2011]. Aggregating recency scores over a time span (10 day-window as in our case) can introduce some documents belonging to different tasks and thus bring more noise to the ranking results. One direction for future work is thus to design a more flexible time window size which adapt to the user's task.

Compared to the sole activity-based ranking methods, adding propagation show clear improvements in prediction, starting from the baseline SUPRA. Bringing semantic relations into the propagation improve even further, producing significantly higher performance for all case of temporal priors. The best performing method, propagation with polynomial de-

Method	S@1	S@10	ARP
MRU	0.162	0.310	76
FMRU	0.131	0.291	87
Ebb	0.213	0.357	65
FEbb	0.193	0.328	70
FPD	0.195	0.331	68
FWei	0.220	0.378	60
PageRank	0.120	0.231	112
SUPRA	$0.320^{\triangle}$	$0.671^{\triangle}$	39
MRU+Prop	$0.353^{\triangle}$	$0.710^{\blacktriangle}$	34
FMRU+Prop	$0.402^{\triangle}$	$0.762^{\blacktriangle}$	30
Ebb+Prop	$0.416^{\triangle}$	$0.733^{\triangle}$	42
FEbb+Prop	$0.452^{\blacktriangle}$	$0.780^{\blacktriangle}$	25
FPD+Prop	$\mathbf{0.512}^{\blacktriangle}$	$\mathbf{0.818}^{\blacktriangle}$	**20**
FWei+Prop	$0.430^{\triangle}$	$0.750^{\blacktriangle}$	40

Table 15: Results on the revisit prediction task. The upper part reports baseline results, the lower part reports results of the proposed system. Symbol $^{\triangle}$ confirms significance against the baseline MRU, and $^{\blacktriangle}$ confirms both significance against the baselines MRU and SUPRA

	Dataset Person				Dataset Collaboration			
Method	P@1	P@10	NDCG@10	MAP	P@1	P@10	NDCG@10	MAP
MRU	0.365	0.283	0.219	0.207	0.461	0.375	0.285	0.267
FMRU	0.329	0.307	0.221	0.213	0.457	0.346	0.271	0.258
Ebb	0.407	0.350	0.258	0.218	0.507	0.392	0.287	0.256
FEbb	0.391	0.292	0.217	0.213	0.493	0.357	0.275	0.260
FPD	0.382	0.290	0.214	0.220	0.480	0.400	0.301	0.288
FWei	0.443	0.402	0.324	0.293	0.552	0.424	0.319	0.290
PageRank	0.318	0.251	0.195	0.164	0.388	0.325	0.195	0.204
SUPRA$^{\triangle}$	0.547	0.502	0.426	0.389	0.590	0.469	0.345	0.333
MRU+Prop$^{\triangle}$	0.518	0.456	0.358	0.333	0.561	0.448	0.334	0.340
FMRU+Prop$^{\triangle}$	0.592	0.511	0.431	0.366	0.630	0.493	0.400	0.361
Ebb+Prop$^{\triangle}$	0.615	0.529	0.503	0.481	0.752	0.642	0.501	0.476
FEbb+Prop$^{\blacktriangle}$	**0.728**	0.621	**0.556**	**0.540**	**0.821**	**0.679**	**0.528**	**0.519**
FPD+Prop$^{\blacktriangle}$	0.710	**0.635**	0.523	0.510	0.780	0.667	0.500	0.482
FWei+Prop	0.678	0.575	0.521	0.478	0.715	0.634	0.479	0.460

Table 16: Performances of ranking methods in the user study. Symbols $^{\triangle}$,$^{\blacktriangle}$ indicate the significance test in all scores of the method against MRU ($p < 0.05$) and SUPRA ($p < 0.01$) respectively.

cay prior, improves the results by 60% as compared to SUPRA. In addition, in contrast to the observed trend in the activity-based ranking, here the combination of frequency and recency with the propagation actually produces better results than the only combination between recency and the propagation.

User-perceived Evaluation

We next aim to evaluate the effectiveness of our proposed system with respect to the user perception and appreciation. We do this by simulating the way users re-access and re-assess the documents in their collections. The experiment is set up the same way as reported in in [Kanhabua et al., 2015], Section 6.1.2. In the dataset Person, each assessor chose 4 weeks to evaluate . For the dataset Collaboration, 2 assessors are asked to choose 3 weeks per each, all are related to joint events they participate in. The activity history is constructed according to this pair of users. The ranking methods are configured to provide the ranks of document with respect to the same time step of the user's evaluations.

Result. The results are summarized in Table **??** for each dataset, as measured as precision, NDCG and MAP scores. The same trend as the prediction task can be observed here: The activity-based ranking methods perform better than PageRank but worse than SUPRA and our propagation variants. Similarly, the frequency-based functions perform worse than the recency ones as isolated methods, but improve the results when combining with the propagation. All propagation methods except the MRU prior-based give higher results than SUPRA. In addition, compared to the prediction task, the performance of all methods in the user-perceived study are slightly higher. This suggests that many documents, although not accessed subsequently, are still deemed "important" to the user.

In conclusion, based on the idea of managed forgetting, our framework unifies evidences from activity logs and semantic relations in a principled way for computing the memory buoyancy of resources. In our method we employ machine learning techniques that automatically learn from the user access history without manual supervision efforts. Our experiments with two real-world datasets have shown that incorporating the importance propagation via semantic relations between resource significantly improves the performance of the method.

6.2 Preservation Value Calculation in the Semantic Desktop (Pilot II)

The Semantic Desktop (SD) is a powerful approach to support organizational as well as personal knowledge management. In ForgetIT, the DFKI developed SD-based pilots (Personal Preservation Pilots I and II in deliverables D9.3 [Maus et al., 2014] and D9.4 [Maus et al., 2015b], respectively) that mainly focus on the latter.

One of the SD's cornerstones is the Personal Information Model (PIMO) which serves as the basis for knowledge representation and provides a common vocabulary across different applications (see also [ForgetIT, 2014]). A PIMO consists of *concepts* (called "things" such as specific topics, projects, persons, tasks, calendar events, ...), *associations* between them (persons are *member of* projects, a task *has topic* SD, ...), and finally, *associated resources* (documents, e-mails, web pages, pictures, ...) [Maus et al., 2013]. In the previous section, we have discussed how such semantic information benefits the memory buoyancy calculation in the SD. In this section, we will provide more details about the

preservation value (PV) calculation. This component has been implemented in Preservation Pilot II. Due to the diversity of situations, we implemented one version for the final evaluation of the Personal Preservation Application Scenario (also referred to as *WP9 evaluation*, see [Maus et al., 2015b] and [Maus et al., 2016]) and an extended one for the PIMO used at the DFKI's knowledge management department. Both versions are technically identical for the most part, but differ in the number of individual evidences they exploit as well as their weighting, and the number of customization options offered to the user. While version-specific details will later be explained in their respective subsections, we first focus on common aspects.

Basically, our algorithm tries to predict a resource's preservation worthiness by taking several aspects into account. It therefore evaluates evidence factors belonging to the six PV dimensions defined for ForgetIT in [Kanhabua et al., 2015]: investment, gravity, social graph, popularity, coverage and quality.

Compared to the new dimensions listed in section 2.2, we initially already addressed the **content type** dimension in **gravity** as a set of heuristics to cover the semantic types of resources (e.g., e-mail vs. contract). Furthermore, we dropped **time** as not sufficiently relevant for the personal preservation scenario as also other dimensions cover time-related aspects (e.g., repeating usage over time periods, see below).

To determine a user's investment spent on a certain resource, for example, the algorithm evaluates the number of annotations, the length of its wiki text and the number of modifications to the resource. The sketch of the algorithm is as follows:

- Suppose a resource only has low evidence values for each dimension, then the sum of these evidences should also be quite low.
- An exceptionally high value in one dimension, e.g., a user spent a lot of investment on a resource, should definitely lead to a high PV, no matter how high or low the other evidences are. (One extraordinary high value can "pull up" the overall score.)
- If a resource only got scores that are slightly above average but this is true for most of the dimensions, then its PV should be relatively high. (The individual scores should sum up ("escalate") to a rather high value.)
- Resources having a combination of mostly low and some average values should only have a low to average PV.

More conceptional details about the different PV dimensions as well as the evidence factors can be found in [Maus et al., 2015b].

After gathering a value for every factor, which also includes a normalization step (details on this will follow later), they are added according to their belonging to one of the six PV dimensions. To combine evidence factors in a way that fits the requirements described before, we use an approach proposed in [Schwarz, 2010] that is based on the *Dempster-Shafer Theory of Evidence* [Gordon and Shortliffe, 1984, Yager and Liu, 2008]. According to this approach, two evidence scores v and w ($v, w \in$ [0,1]) are added as:

$$v' = v \oplus w := 1 - (1 - v) \cdot (1 - w)$$

The $\oplus$-sum of 0.6 and 0.7 would be 0.88, for example.
After adding the different factors for each dimension, the resulting combined evidences are weighted and summarized using the same approach. Next, the resources are assigned to one of the preservation value categories *gold*, *silver*, *bronze* or *wood* (as defined in [Kanhabua et al., 2015]) by applying the respective thresholds.

Which evidence factors are actually evaluated and how they are weighted is influenced by the preservation strategy selected by the user. These strategies are derived from the four personas identified in [Wolters et al., 2015] and D9.4 [Maus et al., 2015b], respectively.

6.2.1 Preservation Value Calculation used in WP9 Evaluation

The WP9 evaluation scenario is quite different from the one at DFKI. Since participants have not used the PIMO before and will actually only use it for some hours (in the evaluation sessions) or at most a few days (compared to months or years in the case of DFKI users), their interaction with it will be limited to a reduced scope of actions (or system features, respectively).

This problem is partly compensated by an initial interview conducted by the experimenters. Participants answer questions about their attitudes and habits towards photo preservation and also give a bit of a background of their lives by talking about the photos they brought for the study. By answering the questions of the ForgetIT survey, the participants' persona can be identified. This persona is then chosen by the experimenter when setting the preservation strategy. Finally, the experimenter will also populate a user's PIMO with some of the concepts (hobbies, activities, places, etc.) mentioned by the participant.

In contrast to the DFKI scenario, the number of possible interactions in the study is too low to perform a rather extensive parameter tuning of the PV algorithm in short time. To address this challenge, we implemented a generator component that produces user metadata according to given parameters:

- The main parameters are the number of photos uploaded by the user and the amount of them being annotated.
- If they are annotated, how long should the text be and how many things (concepts) of the PIMO are mentioned in this text (semantic interconnection).
- Was the text written in a single action or did the user come back to a resource at a later time to correct or extend the annotation.
- Other parameters are image quality scores, photo ratings or the number of rating actions performed (maybe a user re-adjusted the rating at a later time).
- The generator allows using a uniform or linear distribution over each item.

- If the linear one is chosen, there is an additional parameter to set whether lower or higher values of a given range should be preferred.
- For some items the distribution is given by using discrete classes. In the case of image ratings, for example, this could be 10% *favorites*, 60% *liked*, 25% *disliked* and 5% *trashed* photos.

Since the PV calculation only operates on metadata, it is transparent for the algorithm whether it runs on real-world data collected from the interaction with existing photos uploaded to the system or artificial metadata generated according to given constraints.

For the actual PV calculation we use the following evidence factors:

- *investment*: For investment we take the number of annotations, the wiki text length and the number of two kinds of user actions, namely wiki text writing and image rating, into account. Higher values lead to a higher PV. Beside counting the number of rating actions, the actual state whether a resource has been rated or not separately triggers a certain bonus.
- *gravity*: Since we had only a single resource type (images), we omitted gravity here.
- *social graph*: For every person used by the participant, the algorithm checks whether this person can be found in available linked open data sources like DBpedia or Freebase. If this is not the case, we assume that he or she is personally known by the user.[14] Thus, a certain bonus is added to the resource.
- *popularity*: We use the image rating here. Higher ratings lead to higher PV scores.
- *coverage*: Concerning coverage we ensure that at least one photo of each collection gets the highest possible PV (in order to trigger its preservation).
- *quality*: In this dimension we use the image quality. The higher an image's quality score, the higher its PV.

Each evidence factor evaluation should lead to a value in $[0, 1]$. Some items have values in this range already, e.g., user rating classes ranging from *favorite* (1) to *trash* (0). To others, like the length of a wiki text, we apply the following normalization function:

$$n_1(x) := \begin{cases} 1 - \frac{1}{ln(x+e)} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

To additionally cope with the formerly described problem of varying user behavior (generally short vs. long annotations, few vs. most pictures highly rated, etc.), the resulting values are further normalized in a subsequent step. All values in a certain category, e.g., the wiki text length, are divided by their maximum value in this category (in this case the

[14] This is a good example to see the challenges btw. the expected evaluation scenario and the existing, live DFKI PIMO with currently over 450 instances of persons (covering users, colleagues, partners, etc.) allowing a much richer differentiation.

value calculated for the longest text). According to the PV dimensions explained earlier the different factors are then added in order to get combined evidences for investment (e_I), social graph (e_S), popularity (e_P), and quality (e_Q). Let us consider the calculation of e_I as an example. Given the normalized values for the number of annotations (n_A), the wiki text length (n_T), the number of writing actions (n_W) and the number of rating actions (n_R) as well as their respective weights w_A, w_T, w_W and w_R, the combined evidences for investment e_I are defined as follows:

$$e_I(n_A, n_T, n_W, n_R) := w_A \cdot n_A \oplus w_T \cdot n_T \oplus w_W \cdot n_W \oplus w_R \cdot n_R$$

The other combined evidences e_S, e_P and e_Q are calculated analogously. Next, these combined values are summed up:

$$c(e_I, e_S, e_P, e_Q) := 1.0 \cdot e_I \oplus 0.5 \cdot e_S \oplus 0.75 \cdot e_P \oplus 0.5 \cdot e_Q$$

$$e_I, e_S, e_P, e_Q \in [0, 1]$$

$$f(e_I, e_S, e_P, e_Q) := 0.5 \cdot e_I \oplus 0.5 \cdot e_S \oplus 0.75 \cdot e_P \oplus 0.75 \cdot e_Q$$

Basically, for a given resource x, PV(x) is equal to $c(...)$ for *curators* and to $f(...)$ for *filers*. The *safe* variants of these profiles require another process step to ensure *coverage*. For photo collections this means that all photos having the highest PV in their respective collection will receive a preservation value of 1, which is the highest value possible. The PV calculated in a first pass is then overwritten accordingly. In a final step, our algorithm assigns each resource to its matching preservation class by applying the respective thresholds. For the sake of simplicity, only the *gold* and *wood* class were used in the evaluation. Thus, the only relevant threshold to apply was 0.8, so that each resource was either gold (i.e. preserved, PV ≥ 0.8) or wood (i.e. unpreserved, PV < 0.8).

Participants could inspect the algorithm's suggestions by viewing their "time capsule", which basically is just a two-columned table showing the preserved images of each collection on the left and the unpreserved ones on the right-hand side. If they disagreed with the algorithm's classification, they could manually move each image to the respective other side. These actions were logged so that we can later learn from them (together with the participants' statements in a debriefing interview) and improve our algorithms in possible future versions. Further details about the WP9 evaluation can be found in D9.5.

6.2.2 Preservation Value Calculation used at DFKI

Generally speaking, the DFKI version of our algorithm is an extension of the initial version used for the WP9 evaluation. Therefore, we only focus on the additional aspects in this section. For general aspects and those that both versions have in common we kindly refer the reader to the previous sections.

In contrast to the first scenario, DFKI's Knowledge Management department's PIMO contains – depending on the user – data from several months up to several years. Additionally, the scope of possible system interactions (or used system features, respectively) is much higher. On the one hand, this means that there is much more data available to base our

preservation suggestions on. But on the other hand, fine-tuning of the algorithm is more difficult due to the increased number of parameters.

First, the *cold start problem* (see last section) is already solved for most users at DFKI, since they already used the PIMO for a longer time. Second, DFKI users may choose their preservation strategy on their own. They may use the four previously introduced strategies, such as *Safe Curator*, *File & Forget*, etc., as presets, but may additionally check or uncheck individual items (heuristics or rules) in the different dimensions.

It is therefore possible to set the algorithm to measure *investment* only by the number of annotations, disregarding the length of a resource's wiki text and other factors, for example. Relevant evidence factors are:

- *investment*: Same as PV calculation for WP9 evaluation.
- *gravity*: There are several factors indicating a gravity. First, there is a resource's *connectivity*, i.e. the number of connections to other resources (higher connectivity leads to greater PV). Second, we evaluate the *closeness* to another resource we consider to have a certain importance, for example a certain project or event. In addition, the basic gravity value is higher for some resource types. For example, many people are flooded with hundreds of emails every day, so an email should not be declared to be very important per se. For projects we additionally check how many persons are involved in it. More persons imply a higher PV.
- *social graph*: In the DFKI scenario, there are basically two kinds of persons. Those that are also PIMO users and the rest. If, for example, a PIMO user is on a photo, then this photo gets a higher PV on the assumption for the group PIMO, that this person, as a colleague, is more related to the user than an arbitrarily other person. Things presenting a person get a higher PV if the respective person is involved in many projects.
- *popularity*: For popularity we use ratings (in the case of images) as well as the number of views of a resource. The latter was implemented after the WP9 evaluation.
- *coverage*: Same as PV calculation for WP9 evaluation.
- *quality*: The same is true for quality. Like in the evaluation scenario, we only include image quality.

The summarization of the different factors to combined evidences is done analogously to the first scenario, except for the normalization. The DFKI's PV algorithm uses the following functions:

$$n_2(x) := max(\ a \cdot [1 - (x + c)^{-d}],\ 1\ ) \qquad (0 \leq d \leq 1)$$

$$n_3(x) := max(\ a \cdot [1 - \frac{1}{log_b(x + c)}],\ 1\ )$$

Please note that we omitted stating different cases for the domain of x for the sake of readability. If a preservation strategy preset is chosen, the combined evidences for investment (e_I), gravity (e_G), social graph (e_S), popularity (e_P) and quality (e_Q) are summed up as follows:

$$c(e_I, e_S, e_P, e_Q) := 0.8 \cdot e_I \oplus 0.8 \cdot e_G \oplus 0.65 \cdot e_S \oplus 0.5 \cdot e_P \oplus 0.5 \cdot e_Q$$

$$e_I, e_S, e_P, e_Q \in [0, 1]$$

$$f(e_I, e_S, e_P, e_Q) := 0.5 \cdot e_I \oplus 0.5 \cdot e_G \oplus 0.65 \cdot e_S \oplus 0.8 \cdot e_P \oplus 0.8 \cdot e_Q$$

Like in the first scenario, the PV for a given resource is equal to $c(...)$ for *curators* and to $f(...)$ for *filers*. The *safe* variants may also overwrite the PV to ensure a certain *coverage*.

Unlike the evaluation setting, due to the richer material and evidences in the DFKI PIMO, we use all PV categories for classifying a thing according to its PV and the respective threshold of the category. This allows for a more detailed preservation strategy setting allowing different preservation levels for the categories (see preservation strategy in D9.4).

To conclude this section, the PV algorithm achieved good results in the WP9 evaluation (for details please see D9.5 [Maus et al., 2016]). Additionally, the results we observed at DFKI were also promising, so that putting more effort into improving these algorithms as well as doing further research in this area would be justified.

7 Policy-based Preservation Framework

In deliverable [Kanhabua et al., 2015], we have discussed in details the development of our policy-based preservation framework, which is based on the Business Rule Management System (BRMS) Drools. It enables the user of the PoF framework to customize their preservation strategy. The traditional BRMS targets enterprise scenarios and needs to be adapted to ForgetIT preservation scenarios. In this deliverable, we discuss two additional aspects of the policy-based framework. The first aspect relates to how to make the Policy-based more user-friendly, especially to users without background on business rules. The second aspect addresses the problem of uncertainty attached to the policies, which was also a point raised in the review recommendations of the second ForgetIT review. The uncertainty helps to relax the reasoning process in the policy framework, e.g., when dealing with consistency and conflict resolution over the complex set of enterprise policies. Studying a full-fledged uncertain policy resolution system is not the focus in this deliverable, instead we discuss only the relevant concepts, and how this can be adopted into the ForgetIT policy framework.

7.1 User Preference Acceptor and Translator

The problem with BRMS is that users must acquire some rule specification languages in order to be able to define policies to the system. For instance, the Drools Rule Language (DRL) looks as shown for the example rule in Figure 11.

```
Listing 1.
rule "scheduled task document 1.2 1"
when
    exists
        a : rules\_document(correctedPV < 4 and
                            type != "officialDocument" and
                            type != "adminReport")
then
    a.setCorrectedPV("3")
```

Figure 11: Example Drool Rules Language

The syntax is very verbose and error-prone. In ForgetIT, we have designed another interface to ease the selection of rules for basic users. The idea is that the rules are designed by a rule expert together with an *interpretation* of such rules in natural language. The interface shown to the users consists of different questions, grouped in a set of scenarios. Each question has multiple options, each of which corresponds to a rule that specifying users' preferences in preservation for the given scenario.

In addition, to make it easier for users, the default options are provided for each question, matching to the user profile, which is collected as follows. Figure 12 shows "the Sign Up"

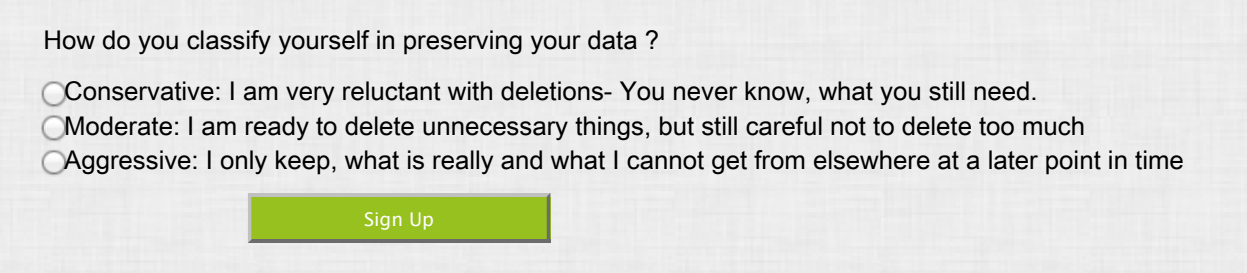

Figure 12: User's general preferences of preservation are asked at Profile Creation page

page, where each user in addition to other information provides information about her own "preservation type". This is done by enabling the users to choose their general attitude towards preservation, as "conservative", "moderate" or "aggressive". Conservative people rarely delete their contents, while moderate people occasionally delete contents they know they no longer need, but are cautious, when they do this. Aggressive people are the most relaxed in removing any unnecessary documents from their systems. Based on this simple claim, the default options are generated for each questions about preservation preferences (see Figure 13).

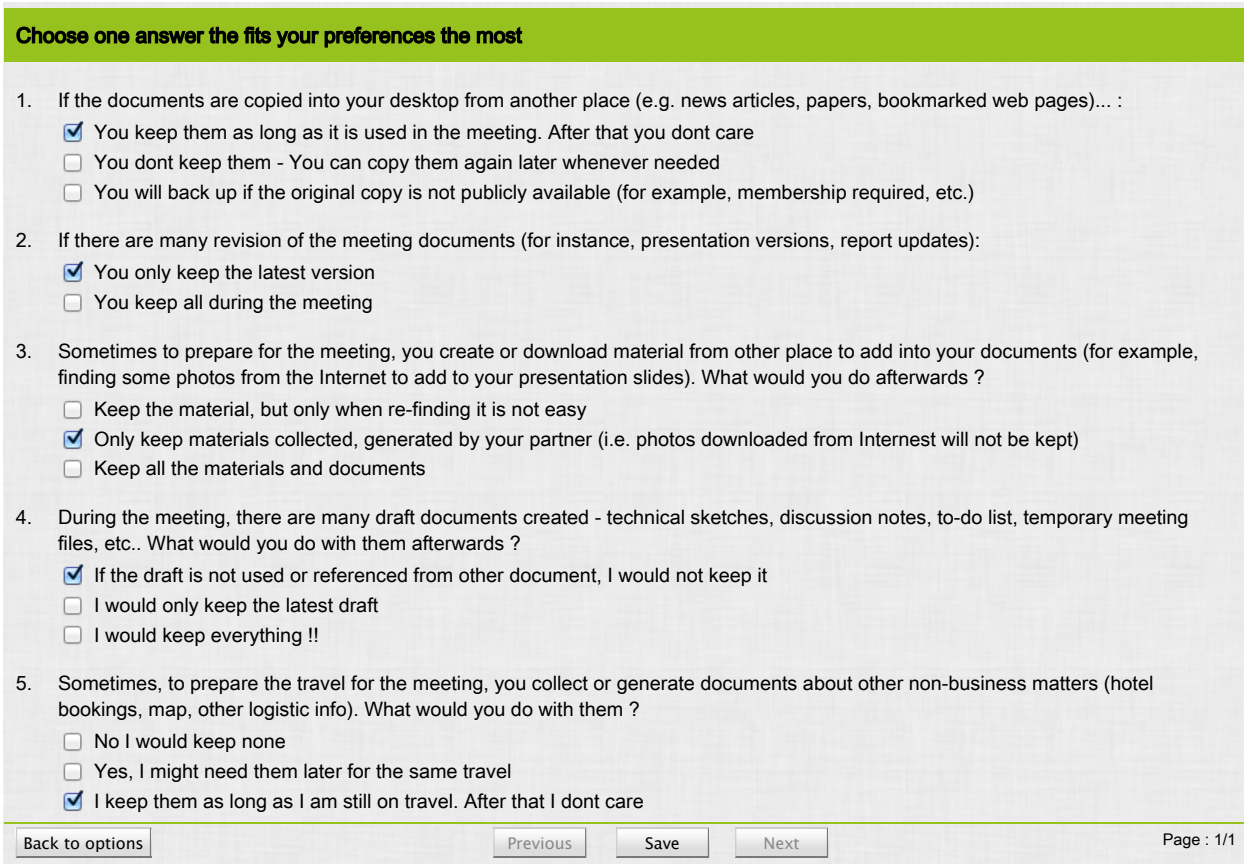

Figure 13: Rules are translated into natural language questions and options. Default options are inferred from the user's general preferences

7.2 Discussion on Uncertainty

In many cases, introducing uncertainty by combining hard logic rules with probability can improve the system performance [Sick, 2002, Kamiya et al., 2005]. In ForgetIT, although we did not yet consider implementation of uncertain rules (soft rules) into our policy framework, we took a literature study to investigate the potential and possibility of integration such approaches into our framework. This section reports our study, and discusses potential directions for an extended framework.

7.2.1 Uncertainty in RETE network

RETE Algorithm Revisit. Our policy-based preservation framework heavily relies on Drools Expert, a rule engine behind implementing the RETE algorithm [Forgy, 1990]. To study the possibility to equip uncertainty to Drools, we first revisit the RETE algorithm foundation, and study to which extent it supports the uncertain inference process. RETE is an algorithm for efficient reasoning over the production rules, determining whether a new rule is allowed or discarded. RETE increases the speed by caching and indexing temporary facts in a *Working Memory*. RETE iteratively asserts facts into the working memory, constructing α-memories, which are assertions of individual facts, and β-memories, which are assertions of facts joined from the individual ones. When constructed with constraints and rules, such memories become the leaves of α- and β-networks, and together these form a *RETE network*. Different underlying data structures such as Hash tables or priority queues can be used to implement the operations on the RETE networks.

Upon the insertion of a new rule, RETE evaluates the rules over the RETE network, propagates the new rule through nodes of the α- and β-networks. At each node, it is combined with existing rules and assertions of existing facts, and performs one of the three operations [Sottara et al., 2010]:

Pass The rule is forwarded with the constraint evaluation result (*true* or *false*).

Hold The rule is held within the node until the evaluation returns a different value.

Drop The rule is discarded.

Extended RETE network. A traditional RETE network only accepts boolean constraints, rules with true constraint are passed and false ones are dropped. [Sottara et al., 2010] proposed extending this architecture by plugging to each rule a new attribute called **degree**, indicating the probability that a property is true in a boolean sense. Note that the degree is not necessarily a numerical value, it can be a range as in the case of Belief Logic Programming (discussed below), or can be a complex object. The extended RETE network consists of an additional component called γ*-network*, where the nodes (called γ-memories) are the evaluation of the nodes in α- and β-networks. The evaluation in each γ-memory node is done by applying a *combination function* (discussed below) on a given α- or β-memory node, generating the values of the degree of the node. The algorithm for

performing the three operations (pass, hold, drop) on the γ-network is discussed in more details in [Sottara et al., 2010].

7.2.2 Belief Logic Programming

Besides proposing an extended RETE network with the introduction of degrees and γ-memories, Sottara et al. also discussed a number of variants for the degree representation. One variant that we found provides a good trade-off between expressivity and complexity is using a numerical range $[v, w]$ to specify the lower and upper bounds for the belief associated with the rule. This representation indeed has been proposed by [Wan and Kifer, 2009] in a theory called *Belief Logic Programming* (BLP). In this deliverable, we choose to study BLP and see how it can be applied to the Drools framework.

Formally, a BLP is a set of **annotated rules**. Each annotated rule has the format: $[v, w]X :- Body$, where X is a positive atom and $Body$ is a boolean combination of atoms. The values $1 \leq v \leq w \leq 1$ specify the degree of the rule, where v specifies the probability that X has true, and $1 - w$ specifies that probability that X is false. In other words, if $Body$ is true, then the rule supports X with the probability v, and supports $\overline{X}$ with the probability $1 - w$. If the $Body$ is a true assertion, then the annotated rule is called an **annotated fact**. Figure 14 illustrates the rule in Listing 11 as in BLP syntax.

```
Listing 2.
rule "scheduled task document 1.2 1"
degree "[0.6,0.8]"
when
    ....
then
    a.setCorrectedPV("3")
```

Figure 14: The rule in Listing 11 in Belief Logic Programming

When combining input atoms, the degree of the output atom is specified via **combination functions**. Formally, let D be the set of all sub-intervals of $[0, 1]$, a function $\Phi : D \times D \rightarrow D$ is called a combination function if it is associative and commutative. These associativity and commutativity properties make it easy to extend a combination function to three or more arguments, and the order of the arguments are immaterial.

[Wan and Kifer, 2009] shows that Belief Logic Programming is a specific case of the **Dempster-Shafer's theory** [Dempster, 1967], where the combination functions are the special forms of Dempster's belief functions. The authors introduced the following three combination functions:

1. Maximum: $\Phi^{max}([v_1, w_1], [v_2, w_2]) = [max(v_1, w_1), max(v_2, w_2)]$

2. Minimum: $\Phi^{min}([v_1, w_1], [v_2, w_2]) = [min(v_1, w_1), min(v_2, w_2)]$

3. Dempster's Combination: $\Phi^{DS}([v_1, w_1], [v_2, w_2]) = \begin{cases} [0,1] & v_1 = w_1 = 0, v_2 = w_2 = 1 \\ [v, w] & otherwise \end{cases}$

where $v = \frac{v_1 w_2 + v_2 w_1 - v_1 v_2}{K}, w = \frac{w_1 w_2}{K}, K = 1 + v_1 w_2 + v_2 w_1 - v_1 - v_2$

Figure 15 illustrates the Dempster's Combination function with three annotated rules and three annotated facts.

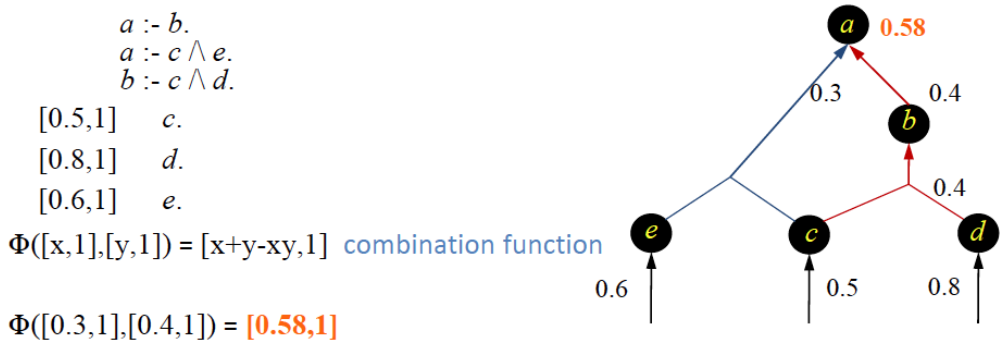

Figure 15: Example Dempster's Combination function on a BLP rule set

7.2.3 Applicability in Drools

Given one specific BLP combination rule, and given the degree of annotated facts, it is trivial to implement the algorithm of extended RETE network described in [Sottara et al., 2010]. Looking deeper into architecture of Drools[15], we found out that it is also possible to plug such BLP components into the Drools Expert engine, via Belief Systems APIs. There are three main interfaces provided:

1. **ModeAssertion** This is the interface to evaluate an abstract node in the RETE network (α-memories, β-memories). In the extended RETE network, we need to specify an additional class, `BLPMode`, to represent the γ-memory, implementing this interface. Among the extra properties of $BLPMode$ are the fields specifying the bouonds of the node's range degree.

2. **BeliefSet** This defines the logical insertions in Drools, or the constraint evaluation in the RETE network. Similarly to ModeAssertion, we need to provide an implementation class `BLPBeliefSet` with additional properties to hold the corresponding degree value. Here the degree value is calculated by applying a specific BLP combination function discussed above.

3. **BeliefSystem** This interface is where we implement the operations on the extended RETE network, by the algorithms described in [Sottara et al., 2010].

[15] Open sourced at `https://github.com/droolsjbpm/drools/tree/master/drools-core/src/main/java/org/drools/core`

In addition, while the system is pluggable, the registration process is currently hard coded into an enum `org.drools.core.BeliefSystemType`, so we need to add there one value referring to the BLP component.

8 Conclusions

8.1 Summary

This deliverable describes our continuous work on managed forgetting including methods for preservation value assessment for different types of content and for different scenarios as well as applications and extensions of the policy framework. We investigate criteria for information value assessment and present our anticipated methodological insights on this issue relating it to the concept of appraisal. In addition, we propose methods to address preservation value assessment in multiple specific scenarios, including preservation value for images, preservation value for text, as well as preservation value for social media. Furthermore, we present two applications related to automatically computing the memory buoyancy of a document with respect to the user attention, and support personal knowledge management via estimating preservation value in the Semantic Desktop. Finally, we address the policy framework and report on a literature study on uncertainty.

8.2 Assessment of Performance Indicators

For progress assessment, we consider the Success Indicators which have been defined in the Description of Work and which are listed below.

- (1) Conceptual process improves selection and preservation activities.
- (2) Ability to complement human memory and meet human expectation.
- (3) Capability to express required constraints (e.g. legal retention constraints).
- (4) Capability to express strategies required by the ForgetIT applications.
- (5) Degree of implementation and integration of defined concepts (forgetting, strategies, information assessment)
- (6) User satisfaction with managed forgetting applications

The Performance Indicator (1) ”Conceptual process improves selection and preservation activities” has already been an imoprtant topic in the deliverables [Kanhabua et al., 2013, Kanhabua et al., 2014]. In this deliverable, the discussion of the presrevation dimensions in section 2 adds further aspects to the conceptual foundations. The Performance Indicator (2) ”Ability to complement human memory and meet human expectation” has been an important topic of deliverable [Kanhabua et al., 2013]. The questionnaires in the context of photo preservation (see deliverable citeD3.3 and work in WP2), in the context of social media content (see deliverable [Kanhabua et al., 2014]) and in the context of scientific situations (see this deliverable) gave further insights into human expectations towards managed forgetting. The Performance Indicator (3) ”Capability to express required constraints” has been adressed by the policy framework, which eanbles the

customization of managed forgetting strategies (see deliverable [Kanhabua et al., 2015] and this deliverable). The Performance Indicator (4) ”Capability to express strategies required by the ForgetIT applications” has been also been adressed by the policy framework, which eanbles the customization of managed forgetting strategies (see deliverable [Kanhabua et al., 2015] and this deliverable) as well as by the consideration of concrete applications for the managed foregetting approach (see [Kanhabua et al., 2015] and this deliverable). The Performance Indicator (5) Degree of implementation and integration of defined concepts has been adressed by building and evaluationd methods for computing memory buoyancy (see deliverable [Kanhabua et al., 2014, Kanhabua et al., 2015] and for computing preservation value (see [Kanhabua et al., 2015] and this deliverable) and by integrating the Forgettor into the PoF Framework. The Performance Indicator(6) ”User satisfaction with managed forgetting applications” has been implicitely adressed by evaluationg the preservation value computation against the actual user selections and more directly by the evaluations of the memory buoyancy computation for the Semantic desktop [Tran et al., 2016a].

8.3 Lessons Learned

We have now worked for more than three years on the topic of managed forgetting including foundations, expectation as well as methods and components for information value assessment in support of computing presrvation value and memory buoyancy. On our journey to managed forgetting in the ForgetIt project we have learned many lessons on various levels. In the following, the most important lessons are summarized:

- Computing preservation value is more complicated than expected and there is no generic method that fits all situations. To address this problem, we worked on a set of solutions for different scenarios (i.e., preservation value for images, for text, and for social media).

- Personal photos are a promising and well-perceived area for managed forgetting and preservation approaches in the personal setting. On the one site, there is a wide undestanding that for the growing amount of content produced more adavanced methods are required to adequately deal with it on the long run. On the other site, recent improvements in image analysis (deep learning) also open new opportunitie for building effective automated selection processes.

- Interdisciplinary work especially with the cognitive scientists has brought many new insights and ideas into our work. Ideas of complementing human memory, a better undestanding of the forgetting process as wel as an understanding of episodic memory and interferences in episodic memory have triggered various of the research ideas followed up in WP3.

- The creation of immediate benefit is crucial for the acceptance of a preservation solution. Long-term benefits as they result from presrvation are not necessarily perceived so strongly.

- The full evaluation of managed forgetting methods and expecially preservation value computation is very dificult due to the inherently long validation delay. So far we mainly evaluate, what peolple think they will still want to have on the long run. For full evaluation a long-term study over several years if not decades would be required.

8.4 Vision for the Future

The ForgetIT project and within the project especially WP3 worked on exploring the area of automatically selecting content of current importance (memory buoyancy) as well as content of long-term importance (preservation value). In both areas considerable progress has been made delivering first viable methods for specific domains such as photo selection, the selection of personal content in more general and for the selection of social media content. Clearly more exciting research is still required for further improving those methods and for opening up further application cases.

During the project we also experienced that the idea of managed forgetting especially for the area of dealing with photos was very well perceived. In the CeBIT 2016, for example, where we presented a photo selection application developed in ForgetIT we received a lot of positive feedback and interest from individuals as well as from companies, who clearly saw the value of such a technology.

The vision is to further explore this area and to come to a situation, where random forgetting of digital content is replaced by user-friendly, semi-automatic forms of managed forgetting, which are for example offered as part of a personal preservation service (see also deliverable [Akşener et al., 2015]).

9 References

[Achanta et al., 2009] Achanta, R., Hemami, S., Estrada, F., and Susstrunk, S. (2009). Frequency-tuned salient region detection. In *IEEE CVPR '09*.

[Akata et al., 2011] Akata, Z., Thurau, C., and Bauckhagee, C. (2011). Non-negative matrix factorization in multimodality data for segmentation and label prediction. In *Proceedings of the 16th Computer Vision Winter Workshop*.

[Akşener et al., 2015] Akşener, F., Sözen, S., Niederée, C., Maus, H., and Nilsson, J. (2015). ForgetIT Deliverable D11.4: Personal Preservation as a Service.

[Andrew and Gao, 2007] Andrew, G. and Gao, J. (2007). Scalable training of l1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

[Backstrom and Leskovec, 2011] Backstrom, L. and Leskovec, J. (2011). Supervised random walks: predicting and recommending links in social networks. In *WSDM*, pages 635–644. ACM.

[Berntsen, 2009] Berntsen, D. (2009). *Involuntary autobiographical memories: An introduction to the unbidden past*. Cambridge University Press.

[Boguraev and Kennedy, 1997] Boguraev, B. and Kennedy, C. (1997). Salience-based content characterisation of text documents. *ACL*.

[Bordes et al., 2005] Bordes, A., Ertekin, S., Weston, J., and Bottou, L. (2005). Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research*.

[Bordino et al., 2013] Bordino, I., Mejova, Y., and Lalmas, M. (2013). Penguins in sweaters, or serendipitous entity search on user-generated content. In *CIKM*.

[Borth et al., 2013] Borth, D., Ji, R., Chen, T., Breuel, T., and Chang, S.-F. (2013). Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM '13*.

[Breiman, 1996] Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[Cauwenberghs and Poggio, 2001] Cauwenberghs, G. and Poggio, T. (2001). Incremental and decremental support vector machine learning. In *Proc. of NIPS*.

[Ceroni et al., 2015a] Ceroni, A., Solachidis, V., Fu, M., Kanhabua, N., Papadopoulou, O., Niederée, C., and Mezaris, V. (2015a). Investigating human behaviors in selecting personal photos to preserve memories. In *Proceedings of the 2015 IEEE International Conference on Multimedia and Expo (ICME) Workshop on Human Memory-Inspired Multimedia Organization and Preservation (HMMP'15)*.

[Ceroni et al., 2015b] Ceroni, A., Solachidis, V., Niederée, C., Papadopoulou, O., Kanhabua, N., and Mezaris, V. (2015b). To keep or not to keep: An expectation–oriented photo selection method for personal photo collections. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR'2015)*.

[Chang et al., 2001] Chang, S. F., Sikora, T., and Puri, A. (2001). Overview of the MPEG-7 standard. *IEEE Trans. Circuits and Systems for Video Technology*, 11(6):688–695.

[Chatzichristofis and Boutalis, 2008] Chatzichristofis, S. A. and Boutalis, Y. S. (2008). Fcth: Fuzzy color and texture histogram - a low level feature for accurate image retrieval. In *Proceedings of the 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, WIAMIS '08, pages 191–196, Washington, DC, USA. IEEE Computer Society.

[Chen, 2005] Chen, Y. (2005). Information valuation for information lifecycle management. In *Proceedings of International Conference on Autonomic Computing*.

[Connor and Spitzer, 2015] Connor, M. and Spitzer, S. (2015 (accessed August 31, 2015)). The places frecency algorithm. `https://developer.mozilla.org/en-US/docs/Mozilla/Tech/Places/Frecency_algorithm`.

[Conway, 2000] Conway, P. (2000). *Overview: Rationale for Digitization and Preservation*. NEDCC.

[Cook, 2005] Cook, T. (2005). Macroappraisal in theory and practice: Origins, characteristics, and implementation in canada, 19502000. *Archival Science*, 5(2-4):101–161.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.*, 20(3):273–297.

[Demartini et al., 2010] Demartini, G., Missen, M. M. S., Blanco, R., and Zaragoza, H. (2010). Taer: time-aware entity retrieval-exploiting the past to find relevant entities in news articles. In *CIKM*.

[Dempster, 1967] Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *The annals of mathematical statistics*, pages 325–339.

[Dunietz and Gillick, 2014] Dunietz, J. and Gillick, D. (2014). A new entity salience task with millions of training examples. *EACL 2014*, page 205.

[Fitchett and Cockburn, 2012] Fitchett, S. and Cockburn, A. (2012). Accessrank: predicting what users will do next. In *SIGCHI*, pages 2239–2242. ACM.

[ForgetIT, 2014] ForgetIT (2014). Deliverable D9.3: Personal Preservation Pilot I: Concise Preserving Personal Desktop.

[Forgy, 1990] Forgy, C. L. (1990). Expert systems. chapter Rete: A Fast Algorithm for the Many Pattern/Many Object Pattern Match Problem, pages 324–341. Los Alamitos, CA, USA.

[Gamon et al., 2013] Gamon, M., Yano, T., Song, X., Apacible, J., and Pantel, P. (2013). Identifying salient entities in web pages. In *CIKM*.

[Ge et al., 2010] Ge, M., Delgado-Battenfeld, C., and Jannach, D. (2010). Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *RecSys*.

[Geng et al., 2008] Geng, X., Liu, T.-Y., Qin, T., Arnold, A., Li, H., and Shum, H.-Y. (2008). Query dependent ranking using k-nearest neighbor. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 115–122, New York, NY, USA. ACM.

[Gordon and Shortliffe, 1984] Gordon, J. and Shortliffe, E. H. (1984). The dempster-shafer theory of evidence. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, 3:832–838.

[Harvey, 2007] Harvey, R. (2007). *Appraisal and Selection*. Digital Curation Center.

[He et al., 2014] He, X., Kan, M.-Y., Xie, P., and Chen, X. (2014). Comment-based multi-view clustering of web 2.0 items. In *Proceedings of the 23rd International Conference on World Wide Web (WWW)*, pages 771–782.

[Huang et al., 2004] Huang, K., Wang, Q., and Wu, Z. (2004). Color image enhancement and evaluation algorithm based on human visual system. In *IEEE ICASSP '04*, volume 3.

[Huiskes and Lew, 2008] Huiskes, M. J. and Lew, M. S. (2008). The mir flickr retrieval evaluation. In *MIR*, pages 39–43, New York, NY, USA. ACM.

[Itten, 1973] Itten, J. (1973). *The art of color : the subjective experience and objective rationale of color*. John Wiley New York.

[John and Langley, 1995] John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proc. of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 338–345. Morgan Kaufmann.

[Kamiya et al., 2005] Kamiya, A., Ovaska, S. J., Roy, R., and Kobayashi, S. (2005). Fusion of soft computing and hard computing for large-scale plants: a general model. *Applied Soft Computing*, 5(3):265–279.

[Kanhabua et al., 2015] Kanhabua, N., Niederée, C., Ceroni, A., Naini, K. D., Kawase, R., Tran, T., Maus, H., and Schwarz, S. (2015). ForgetIT Deliverable D3.3: Strategies and Components for Managed Forgetting - Second Release.

[Kanhabua et al., 2013] Kanhabua, N., Niederée, C., Loggie, R., Tran, T., Djafari-Naini, K., Maus, H., and Schwarz, S. (2013). Deliverable D3.1: Report on Foundations of Managed Forgetting.

[Kanhabua et al., 2014] Kanhabua, N., Niederée, C., Tran, T., Nguyen, T. N., Djafari-Naini, K., Kawase, R., Schwarz, S., and Maus, H. (2014). Deliverable D3.2: Components for Managed Forgetting – First Release.

[Kawase et al., 2011] Kawase, R., Papadakis, G., Herder, E., and Nejdl, W. (2011). Beyond the usual suspects: Context-aware revisitation support. In *Proceedings of the 22Nd ACM Conference on Hypertext and Hypermedia*, HT '11, pages 27–36, New York, NY, USA. ACM.

[Lavoie and Dempsey, 2004] Lavoie, B. and Dempsey, L. (2004). Thirteen ways of looking at...digital preservation. *D-Lib Magazine*, 10(7/8).

[Lee et al., 2007] Lee, H., Battle, A., Raina, R., and Ng, A. Y. (2007). Efficient sparse coding algorithms. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pages 801–808.

[Li et al., 2003] Li, J., Lim, J. H., and Tian, Q. (2003). Automatic summarization for personal digital photos. In *Proceedings of ICICS-PCM '03*.

[Liensberger et al., 2009] Liensberger, C., Stottinger, J., and Kampel, M. (2009). Color-based and context-aware skin detection for online video annotation. In *IEEE MMSP '09*.

[Liu et al., 2013] Liu, J., Wang, C., Gao, J., and Han, J. (2013). Multi-view clustering via joint nonnegative matrix factorization. In *SIAM International Conference on Data Mining (SDM)*, pages 252–260.

[Logie et al., 2014] Logie, R., Niven, E., Wolters, M., and Mayer-Schönberger, V. (2014). ForgetIT Deliverable D2.2: Foundations of Forgetting and Remembering: Preliminary Report.

[Luo and Tang, 2008] Luo, Y. and Tang, X. (2008). Photo and video quality evaluation: Focusing on the subject. In *ECCV '08*.

[Lux and Chatzichristofis, 2008] Lux, M. and Chatzichristofis, S. A. (2008). Lire: lucene image retrieval: an extensible java cbir library. In *Proceedings of the 16th ACM international conference on Multimedia*, MM '08, pages 1085–1088, New York, NY, USA. ACM.

[Machajdik and Hanbury, 2010] Machajdik, J. and Hanbury, A. (2010). Affective image classification using features inspired by psychology and art theory. In *ACM MM'10*.

[Maus et al., 2013] Maus, H., Schwarz, S., and Dengel, A. (2013). Weaving personal knowledge spaces into office applications. In Fathi, M., editor, *Integration of Practice-Oriented Knowledge Technology: Trends and Prospectives*, pages 71–82. Springer.

[Maus et al., 2014] Maus, H., Schwarz, S., Eldesouky, B., Jilek, C., Wolters, M., and Loğoğlu, B. (2014). ForgetIT Deliverable D9.3: Personal Preservation Pilot I: Concise preserving personal desktop.

[Maus et al., 2015a] Maus, H., Schwarz, S., Jilek, C., and Gallo, F. (2015a). ForgetIT Deliverable D9.4: Personal Preservation Pilot II: Concise preserving mobile information assistant.

[Maus et al., 2015b] Maus, H., Schwarz, S., Jilek, C., and Gallo, F. (2015b). ForgetIT Deliverable D9.4: Personal Preservation Pilot II: Concise preserving mobile information assistant.

[Maus et al., 2016] Maus, H., Schwarz, S., Jilek, C., Wolters, M., Rhodes, S., Ceroni, A., and Gür, G. (2016). ForgetIT Deliverable D9.5: Personal Preservation Report.

[Mavridaki and Mezaris, 2015] Mavridaki, E. and Mezaris, V. (2015). A comprehensive aesthetic quality assessment method for natural images using basic rules of photography. In *IEEE ICIP '15*.

[McCreadie et al., 2014] McCreadie, R., Macdonald, C., and Ounis, I. (2014). Incremental update summarization: Adaptive sentence selection based on prevalence and novelty. In *CIKM*.

[Mitra et al., 2008] Mitra, S., Winslett, M., and Hsu, W. W. (2008). Query-based partitioning of documents and indexes for information lifecycle management. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, SIGMOD '08, pages 623–636.

[Naini et al., 2015] Naini, K. D., Altingovde, I. S., Kawase, R., Herder, E., and Niederée, C. (2015). Analyzing and predicting privacy settings in the social web. In *User Modeling, Adaptation and Personalization - 23rd International Conference, UMAP 2015, Dublin, Ireland, June 29 - July 3, 2015. Proceedings*, pages 104–117.

[Obrador et al., 2010] Obrador, P., de Oliveira, R., and Oliver, N. (2010). Supporting personal photo storytelling for social albums. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 561–570, New York, NY, USA. ACM.

[Palpanas et al., 2004] Palpanas, T., Vlachos, M., Keogh, E., Gunopulos, D., and Truppel, W. (2004). Online amnesic approximation of streaming time series. In *Proceedings of the 20th International Conference on Data Engineering*, ICDE '04, pages 338–349.

[Papadakis et al., 2011] Papadakis, G., Kawase, R., Herder, E., and Niederee, C. (2011). A layered approach to revisitation prediction. In *International Conference on Web Engineering (ICWE)*, volume 6757, pages 258–273.

[Papadopoulou et al., 2014] Papadopoulou, O., Mezaris, V., Solachidis, V., Ioannidou, A., Eldesouky, B. B., Maus, H., and Greenwood, M. A. (2014). ForgetIT Deliverable D4.2: Information analysis, consolidation and concentration techniques, and evaluation – First release.

[Peetz and de Rijke, 2013] Peetz, M.-H. and de Rijke, M. (2013). Cognitive temporal document priors. In *Proceedings of the 35th European conference on Advances in Information Retrieval*, ECIR'13, pages 318–330.

[Peetz and De Rijke, 2013] Peetz, M.-H. and De Rijke, M. (2013). Cognitive temporal document priors. In *ECIR*, pages 318–330. Springer.

[Rabbath et al., 2011] Rabbath, M., Sandhaus, P., and Boll, S. (2011). Automatic creation of photo books from stories in social media. *ACM TOMM*.

[Savakis et al., 2000] Savakis, A. E., Etz, S. P., and Loui, A. C. P. (2000). Evaluation of image appeal in consumer photography.

[Schellenberg, 1956] Schellenberg, T. R. (1956). The appraisal of modern records. *Bulletins of the National Archives*, 8:46 pages.

[Schmidt et al., 2007] Schmidt, M., Fung, G., and Rosales, R. (2007). Fast optimization methods for l1 regularization: A comparative study and two new approaches. In *Proceedings of the 18th European Conference on Machine Learning*, ECML '07, pages 286–297, Berlin, Heidelberg. Springer-Verlag.

[Schwarz, 2010] Schwarz, S. (2010). *Context-Awareness and Context-Sensitive Interfaces for Knowledge Work*. PhD thesis, University of Kaiserslautern, Department of Computer Science.

[Seah et al., 2014] Seah, B.-S., Bhowmick, S. S., and Sun, A. (2014). Prism: Concept-preserving social image search results summarization. In *Proceedings of SIGIR '14*.

[Sharma et al., 2005] Sharma, G., Wu, W., and Dalal, E. N. (2005). The CIEDE2000 color-difference formula: implementation notes, supplementary test data, and mathematical observations. *Color research and application*, 30(1).

[Sick, 2002] Sick, B. (2002). Fusion of hard and soft computing techniques in indirect, online tool wear monitoring. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 32(2):80–91.

[Sinha et al., 2011] Sinha, P., Mehrotra, S., and Jain, R. (2011). Summarization of personal photologs using multidimensional content and context. In *Proceedings of ICMR '11*.

[Solachidis et al., 2016] Solachidis, V., Apostolidis, E., Markatopoulou, F., Galanopoulos, D., Tzelepis, C., Arestis-Chartampilas, S., Pournaras, A., Tastzoglou, D., Mezaris, V., Chen, D., Harnik, D., Khaitzin, E., Eldesouky, B., Maus, H., Greenwood, M., and Tan, A. S. (2016). ForgetIT Deliverable D4.4: Information analysis, consolidation and concentration techniques, and evaluation - Final release.

[Solachidis et al., 2015] Solachidis, V., Papadopoulou, O., Apostolidis, K., Ioannidou, A., Mezaris, V., Greenwood, M., and Maus, H. (2015). Deliverable D4.3: Information Analysis, Consolidation and Concentration Techniques, and Evaluation – Second Release.

[Sottara et al., 2010] Sottara, D., Mello, P., and Proctor, M. (2010). A configurable rete-oo engine for reasoning with different types of imperfect information. *Knowledge and Data Engineering, IEEE Transactions on*, 22(11):1535–1548.

[Tamura et al., 1978] Tamura, H., Mori, S., and Yamawaki, T. (1978). Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*.

[Tran et al., 2015a] Tran, N. K., Ceroni, A., Kanhabua, N., and Niederée, C. (2015a). Back to the past: Supporting interpretations of forgotten stories by time-aware re-contextualization. In *Proceedings of International Conference on Web Search and Data Mining*, WSDM '15.

[Tran et al., 2016a] Tran, T., Schwarz, S., Niedere, C., Maus, H., and Kanhabua, N. (2016a). The forgotten needle in my collections: Task-aware ranking of documents in semantic information space. In *Proceedings of the 1st ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR)*. ACM.

[Tran et al., 2016b] Tran, T., Schwarz, S., Niederee, C., Maus, H., and Kanhabua, N. (2016b). The forgotten needle in my collections: Task-aware ranking of documents in semantic information space. In *ACM SIGIR Conference on Computer Human Interaction and Retrieval*. ACM.

[Tran et al., 2015b] Tran, T. A., Niederee, C., Kanhabua, N., Gadiraju, U., and Anand, A. (2015b). Balancing novelty and salience: Adaptive learning to rank entities for timeline summarization of high-impact events. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pages 1201–1210. ACM.

[Valdez and Mehrabian, 1994] Valdez, P. and Mehrabian, A. (1994). Effects of color on emotions. In *Journal of Experimental Psychology*.

[van de Weijer et al., 2007] van de Weijer, J., Schmid, C., and Verbeek, J. (2007). Learning color names from real-world images. In *In IEEE CVPR'07*.

[van den Hoven and Egge, 2014] van den Hoven, E. and Egge, B. (2014). The cue is key - design for real-life remembering. *Zeitschrift für Psychologie.*, 222(2):110–117.

[Walber et al., 2014] Walber, T., Scherp, A., and Staab, S. (2014). Smart photo selection: Interpret gaze as personal interest. In *Proceedings of CHI '14*.

[Wan and Kifer, 2009] Wan, H. and Kifer, M. (2009). Belief logic programming: Uncertainty reasoning with correlation of evidence. In *Logic Programming and Nonmonotonic Reasoning*, pages 316–328. Springer.

[White et al., 2010] White, R. W., Bennett, P. N., and Dumais, S. T. (2010). Predicting short-term interests using activity-based search context. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1009–1018.

[Witten and Frank, 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Wolters et al., 2014] Wolters, M. K., Niven, E., and Logie, R. H. (2014). The art of deleting snapshots. In *Proceedings of CHI EA'14*.

[Wolters et al., 2015] Wolters, M. K., Niven, E., Runardotter, M., Gallo, F., Maus, H., and Logie, R. H. (2015). Personal photo preservation for the smartphone generation. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, Seoul, CHI 2015 Extended Abstracts, Republic of Korea, April 18 - 23, 2015*, pages 1549–1554.

[Wu and Giles, 2013] Wu, Z. and Giles, C. L. (2013). Measuring term informativeness in context. In *NAACL-HLT*.

[Yager and Liu, 2008] Yager, R. R. and Liu, L. (2008). *Classic works of the Dempster-Shafer theory of belief functions*, volume 219. Springer.

[Yeh et al., 2010] Yeh, C.-H., Ho, Y.-C., Barsky, B. A., and Ouhyoung, M. (2010). Personalized photograph ranking and selection system. In *Proceedings of MM '10*.