www.forgetit-project.eu

ForgetIT

Concise Preservation by Combining Managed Forgetting
and Contextualized Remembering

Grant Agreement No. 600826

Deliverable D3.2

Work-package	WP3: Managed Forgetting Methods
Deliverable	D3.2: Components for Managed Forgetting - First Release
Deliverable Leader	Nattiya Kanhabua (LUH)
Quality Assessor	Tero Päivärinta (LTU)
Estimation of PM spent	12
Dissemination level	Public
Delivery date in Annex I	31 January 2014
Actual delivery date	07 February 2014
Revisions	2
Status	Final
Keywords:	Digital Preservation; Dynamic Information Assessment; Time-aware Information Access; Managed Forgetting

Disclaimer

This document contains material, which is under copyright of individual or several ForgetIT consortium parties, and no copying or distributing, in any form or by any means, is allowed without the prior written agreement of the owner of the property rights.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the ForgetIT consortium as a whole, nor individual parties of the ForgetIT consortium warrant that the information contained in this document is suitable for use, nor that the use of the information is free from risk, and accepts no liability for loss or damage suffered by any person using this information.

This document reflects only the authors' view. The European Community is not liable for any use that may be made of the information contained herein.

List of Authors

Partner Acronym	Authors
LUH	Claudia Niederée
LUH	Nattiya Kanhabua
LUH	Tuan Tran
LUH	Tu Ngoc Nguyen
LUH	Kaweh Djafari-Naini
LUH	Ricardo Kawase
DFKI	Sven Schwarz
DFKI	Heiko Maus

Contents

List of Authors **3**

Contents **5**

Executive Summary **7**

1 Introduction **8**

1.1 Deliverable Organization 8

2 ForgetIT Approach to Information Value Assessment **10**

2.1 ForgetIT Model underlying Information Value Assessment 10

2.1.1 Overview 10

2.1.2 Resource Description Framework - RDF 11

2.1.3 Information and Interaction Model 12

2.2 Memory Buoyancy Assessment Approach 15

2.2.1 Resource Re-access Model 16

2.2.2 Building Relation Graph from the PIMO Ontology and Interaction Log 20

2.3 Prototype Components for Information Assessment 20

2.3.1 Prototype Overview 20

2.3.2 RESTful Web Service 22

3 Information Value Assessment: Case Studies **24**

3.1 Effectiveness of Temporal Features 24

3.1.1 Motivation 24

3.1.2 Evaluation 25

3.2 Effectiveness of Social Features 26

3.2.1 Motivation 26

3.2.2 Evaluation 27

3.3 Efficiency of Information Value Assessment 28

3.3.1 Motivation 28
3.3.2 Evaluation 29

4 Complementing Human Memory: Case Studies 30

4.1 How Social Media Complement Human Memory in Public Events: Case Study of Wikipedia and Twitter Trending Topics 30
4.1.1 Motivation 30
4.1.2 Methodology 31
4.1.3 Preliminary Results and Discussion 33
4.2 Understanding Collective Memory in Wikipedia 33
4.2.1 Motivation 33
4.2.2 Forgetting and Memory Catalysts 34

5 Research Plans and Future Activities 37

5.1 Information Assessment in Photo Preservation Scenarios 37
5.1.1 Motivation 37
5.1.2 Methodology and Evaluation Plan 37
5.2 Temporal Summarization of Social Web Content 38
5.2.1 Motivation 38
5.2.2 Methodology and Evaluation Plan 39
5.3 Managed Forgetting Strategies 39

6 Conclusions 40

References 41

Executive summary

The previous deliverable D3.1 addresses the foundations of managed forgetting. More precisely, we outlined the state of the art in and ideas for associating human and digital remembering and forgetting, as well as presented several key research questions for the introduction of the managed forgetting concept into information management. We developed first ideas on how managed forgetting methods can complement human remembering and forgetting processes. In addition, we proposed our first ideas for managed forgetting by presenting a conceptual model and a computational framework. We envisioned that managed forgetting can be regarded as functions of attention and significance dynamics relying on multi-faceted information value assessment.

In this deliverable (D3.2), we first describe models and a framework for information value assessment, which is a core part of managed forgetting. In the first project year, we focus on the assessment of *memory buoyancy*, whereas the *preservation value* will be studied and reported in the next deliverable. Moreover, we explain our research studies for managed forgetting ideas by addressing two main aspects, namely, features for information value assessment, and complementing human memory. Managed forgetting is a novel concept and therefore requires more exploratory research for various aspects of the concept. Thus, we conduct exploratory research in order to gain ideas for a proof-of-concept realization. In detail, we studied relevant features for information value assessment, propose several information value assessment methods as well as present evaluation results in term of effectiveness and efficiency. Our experimental findings have resulted into important insights for the further work on the managed forgetting solutions and demonstrated the feasibility of of the proposed methods and their incorporation into the managed forgetting framework.

1 Introduction

The goal of WP3 is to develop concepts and methods for managed forgetting and to integrate them into the Preserve-or-Forget framework. The methods aim to match human expectations and to complement processes of human forgetting and remembering. The development of managed forgetting methods embraces a conceptual foundation, methods for information value assessment in support of memory buoyancy (short to mid-term importance) and preservation value (long-term importance, and the development of a managed forgetting method, which supports forgetting options as well as forgetting strategies.

In this deliverable (D3.2), we first describe models and an information value assessment framework. In our proposed framework, we define the environment and actions in which forgetting and preservation scenarios and functionality are studied and highlighted. There are three main concepts introduced: *Resources*, *Interactions* and *Human actors*. A resource can be represented by a data object such as a document, image, etc. in an information space, but can also be the human perception of that object. To accommodate this blurred line, we employ concepts in artificial intelligence and Semantic Web, and use ontology from the PIMO semantic desktop system to define the information space. The concept of managed forgetting is relatively new and there is no standard test collections available for proofing the concept. Hence, we opt to perform exploratory research for a realization of certain methods to demonstrate their feasibility for managed forgetting. Given our initial framework of information value assessment, we conduct several studies in more general settings for evaluating the impacts of different features that can be useful the learning phase (for both memory buoyancy and preservation value models). In particular, we are interested in evaluating the effectiveness and efficiency of features for information value assessment. Our experimental findings have demonstrated the feasibility of the proposed methods for incorporating into the managed forgetting framework.

In addition to the studies of information assessment, another aspect related to managed forgetting is to understand to which extent human can remember details or general shape of an event in their real life. Although the sources of such episodic memory vary vastly from individual to individual, in the larger global scale, we believe that there are some common features that govern the human remembering towards public events. Getting insights into such features can greatly help computers to measure preservation value of a digital objects, by associating it with different events. To cope with this problem, we present our preliminary research results on complementing human memory. Finally, we outline planned research activities for further supporting the managed forgetting concept that will be conducted in the next months.

1.1 Deliverable Organization

The detailed organization of the deliverable is outlined below.

- Section 2 describes our proposed framework for information value assessment, which extends the conceptual and computational models of managed forgetting presented in D3.1.
- Section 3 presents research studies on information value assessment in more general settings, and extensive experiments for evaluating the effectiveness and efficiency of our proposed methods.
- Section 4 presents the preliminary results of complementing human memory studies, which include: 1) using Wikipedia to analyse Twitter trending topics, and 2) analysing collective memory in Wikipedia.
- Section 5 outlines our research plan for the next months in WP3 and reports our preliminary research results for further shedding light on the ideas of complementing human memory.
- Section 6 summarizes and concludes the deliverable.

2 ForgetIT Approach to Information Value Assessment

2.1 ForgetIT Model underlying Information Value Assessment

In the previous deliverable D3.1, we have proposed an abstract model of how to integrate managed forgetting process into a personal / organizational information management system. Our abstract model is driven by the idea of relating human memory processes with the organization of digital memory, which is represented by an information space. Inside this space, organizational principles are based on human perception about organizing items (for example, people tend to keep items that are related within a certain task / event / topics closely together), and many functionalities follow the human mental process of organizing or searching particular resources. However, as the aim of the managed forgetting is not only to simulate human brain, but to complement its limitations, there are additional features and functionality required for supporting this goal.

For this purpose, we introduce a model for the information in the information space and for the interactions of an agent with the information space, which provides the basis for information value assessment methods that complement human memory. This section describes first the model and its foundations and subsequently the information value assessment method built on top of this model.

2.1.1 Overview

In our proposed framework, we define the environment and actions in which forgetting and preservation scenarios and functionality are studied and highlighted. There are three main concepts introduced: *Resources*, *Interactions* and *Human actors*. A resource can be represented by a data object such as a document, image, etc. in an information space, but can also be the human perception of that object.

To accommodate this blurred line, we employ concepts in artificial intelligence and Semantic Web, and use ontology to define the information space, where each resource correspond to an *entity*. Each entity is uniquely identified and can have multiple properties, including both properties associated with the corresponding objects (for instance, document size, creation time, etc.) and properties associated with the human interpretation of the objects (for instance, the event or topic by which the human chooses to organize a photo. This can be in a simple form such as user-defined tags for the photos, but can also represented by more advanced concept such as another resource). We use the class **Resource** to group all entities that represent a resource in ForgetIT information space. Similarly, Human actors are also defined as entities, and they are grouped under the special class called **Agent**. Like Resource entities, an Agent entity also has properties to describe the respective human profiles and characteristics (e.g. gender, name, profession, . . .).

One entity can have several relations with other entities, some of which form the **Context** surrounding the entity. For the *Interaction*, in order to facilitate the efficient processing and to cope with the rapid change of human actions and mental processes (adding and re-organize items, looking up old documents, ...), we define a lightweight relational schema called **Action** to model human activities. In the following, we describe our data models in more details.

2.1.2 Resource Description Framework - RDF

Among different ontological knowledge representation, we choose Resource Description Framework (RDF[1]) for its simplicity and flexibility. RDF was designed to describe concepts and meta data across various resources on the Web, and has become a W3C standard for representing knowledge, and for storing and exchanging information in Semantic Web activities. The key point in RDF is that everything (people, cities, artifacts, concepts, etc.) is uniquely represented by a resource, or an entity from the ontological point of view [44]. Typically, an entity is identified by a Uniform Resource Identifier (URI), which is a string following specific syntaxes [5]. For example, `http://www.w3.org/People/Berners-Lee/` is a URI identifying an entity named “Tim Berners-Lee”. A fact in RDF is represented by a triple <*subject*, *predicate*, *object*>, and specifies a relationship between the subject and the object via the property encoded in the predicate. For example, to specify the fact that Tim Berners-Lee is a person, we have the triple as shown in Figure 1:

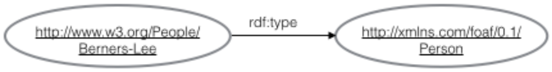

Figure 1: Example RDF triple

Class / **Individual** / **Literal** A resource in RDF corresponds to either a named entity (such as *Berlin*), a relation name (such as *hasCapital* or *wasBornOnDate*), or a literal (such as “30-04-1777”). Similar entities can be grouped into classes. For instance, Gauss and Riemann are grouped into the class person. A class is also an entity. Named entities which are not classes are called individuals. For example, *Germany* is an individual of a class *Country*.

RDFS RDF was extended to Resource Description Framework Schema (RDFS) with new specifications and vocabularies to structure RDF facts in a more expressive way. Among others, some notable improvements are: (1) RDFS allows multiple classes to be grouped further in a super-class by introducing a new relation name *subClassOf*; (2) enabling literals to have certain data types (such as Integer, Date, String, etc.). Further explanation of RDFS goes beyond the scope of this deliverable. More detailed description can be found in [6].

[1] `http://www.w3.org/RDF`

2.1.3 Information and Interaction Model

In ForgetIT, we use RDF to model resources and human actors and their relationships. To resolve the general domain, we use the prefix “forgetit:” which refers to `http://www.forgetit-project.eu/`. To resolve entities and classes used in personal settings, we employ PIMO ontology and introduce new classes under the “forgetit:pimo:” (`http://www.forgetit-project.eu/pimo`) prefix. Similarly, the prefix “forgetit:typo3” (`http://www.forgetit-project.eu/typo3`) is used to resolve the organization settings.

Figure 2 shows the most basic classes in our data model. As in this deliverable, we focus on the personal preservation scenario, we describe here the most relevant classes and relations within the PIMO semantic desktop system. Detailed data model for organization use cases will be described in the subsequent deliverables.

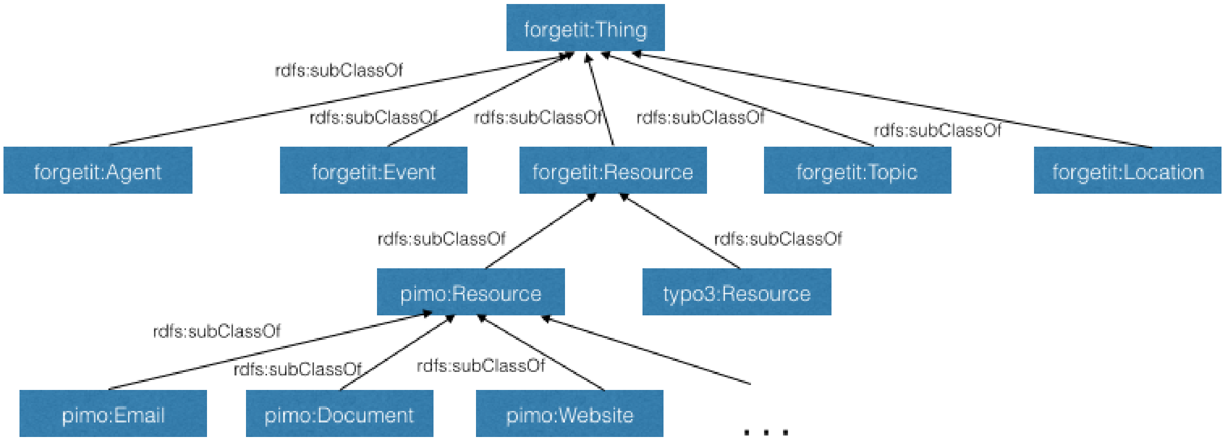

Figure 2: Excerpt of RDF classes in ForgetIT data model

Example RDF facts that describe the information space properties are in Figure 3 (note that here we use labels to identify entities for the sake of brevity and clarity. The real identifiers of entities are resolved via the component *ID Manager*, which is described in work package 8).

Agent An Agent entity represents a person or an organization that interacts with the information space, generate and consume data from the space, and has certain forgetting and preservation need. Agent has not only properties that reflect the personal or organizational profile (e.g. Email address), but also properties that are attached to other types of entities (resources, events, etc.). Note that the actions of the Agent entity performed on a Resource entity (Interactions) are not modeled as part of the ontology, but separately using a relational schema (see below).

Resource A Resource entity encodes the semantic information about a data object in the information space, for instance a stored photo in a personal folder. A resource reflects human perception on the object and not the physical manifestation of itself, thereby helping the interaction between human memory and digital memory (D3.1). For example, a photo can have multiple copies stored in different locations - personal desktop, mobile phone, or in the cloud. However, all these physical files are represented by only a single Resource entity that encodes their semantics.

Event / Topic Event entities reflects the real-world incident that has impact on agents or the resources (for example, a holiday where the agent was on and from which the resource data object is generated.) Topic entities represent the abstract groupings of the resource, and it reveals how the agent see the resources in organized manner. For instance, a research document containing information about human forgetting process can be annotated under the topic “Human Forgetting”, which makes it easier for the agent to retrieve, organize or preserve the document in the future.

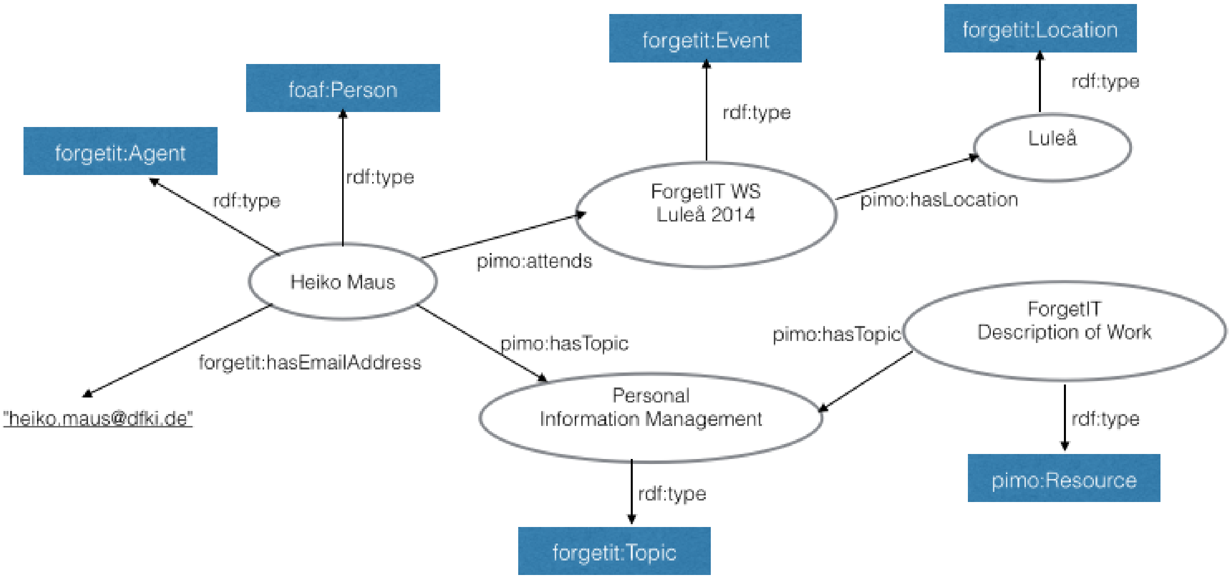

Figure 3: Example of RDF facts in PIMO system

Context

One special concept in ForgetIT technology is context. It reflects the surroundings of an entity, be they the relationships between the entities and the others, or the properties that describe contextual information of the entity itself. In ForgetIT data model, a context is a set of RDF facts centered around an entity (called a *Context Graph* of the entity), where the entities or literals having no direct relation to the entities are omitted. It is still to be researched, if it makes sense to use all types of relations for the construction of this graph and what is the limitation for the distance between the considered entity and the entities in the context. The desired degree of completeness of the context graph that is good enough when the entity information is to be preserved or retrieved will be investigated by the Contextualization / Re-contextualization work in WP6.

In addition, the context is associated with several time literals that encode moments in which an aspect in the context of the entity is captured (for instance, when a person occupation title is changed). This time literal is attached to the context graph by the means of RDF reification [2], i.e. by adding a new RDF triple about time to existing triples to describe the existing triples' time dimension.

Interactions

Interactions refer to the activities of the human actor on the resources within the information space. Interactions consists of a series of atomic *action*, each of which is associated with at least two entities: An agent and a resource. An action has a timestamp property capturing the time point where the interaction is observed, and list of extra information to accommodate in particular scenarios. In ForgetIT, we choose the relational schemas to represent interactions for two reasons. First, there is no reasoning entailed in handling interactions (except when going to quest the detailed information about the resource or the agent), and therefore a full-fledge RDF models is not necessary here. Second, relational schemas can be easily stored in a Relational Database Management System (RDBMS), which supports efficient inserting, updating and retrieving. This is highly desirable in the case when several interactions are observed within a short time span (in a matter of minutes or seconds), and the system needs to responsively capture, store and index the desired actions into a repository.

Table 1 shows an example schema of PIMOAtion, one of the two Action types in ForgetIT, corresponding to the personal and organizational preservation scenarios. Note that in this deliverable, we focus on the actions in PIMO systems; the TYPO3Action relates to the activities within organization scenarios and will be studied in subsequent work. From the implementation point of view, the Action is an abstract class in the managed forgetting component, where some of the operations can be re-used in the PIMOAction and TYPO3Action types. For example, operations on resolving the timestamps of the action is reusable in both PIMOAction and TYPO3Action entries. As shown in the table, the conceptUri refers to the URIs of the resource entity, while the occurenceUri refers to the location of the physical data object (remember resource entity can have multiple physical images). The "Meta-data" attribute is to add the flexibility to the action entries that are captured at the client and sent to the managed forgettor - the client can include some additional information about its environment, about the actions in place to ease the processing at the middle layer.

Timestamp	1336373204762
Application	"MS-PowerPoint"
UserUri	pimo:1327593979868:1
conceptUri	pimo:1332497855250:7
occurenceUri	file:C:/Users/Heiko Maus/Desktop/ForgetIT-DFKI Vision.pptx
Action	access
Meta-data	"actionObjectType":"File"

Table 1: Example of one PIMOAction entry

[2] http://en.wikipedia.org/wiki/Reification_(computer_science)

2.2 Memory Buoyancy Assessment Approach

In this deliverable, we aim to devise a prototype of the information memory buoyancy assessor, one of the core part in the managed forgetting process. Recall that memory buoyancy is a proposed concept in ForgetIT which indicates the short-term to mid-term importance of a digital object reflecting some type of "closeness" to current human memory processes and current activities.

Due to the idea of complementing human memory, memory buoyancy is related in a non-trivial way to the interaction with resources and to the degree to which memories and an image of a digital object in human brain fade over time, either as a result of the decay and interference in human memory, or because of other factors intervening the remembering or recall process. On the one hand, it is important to identify the objects that are currently important or will be important in the near future, i.e., which have a high probability to be (re-)accessed. On the other hand, for complementing human memory, it is also important to consider both a) how easy is it to find the identified important resources (access effort) and b) how probable is it that the actor needs to rehearse the information from the resource (vs. he still remembers the content and, thus, will not re-access). For identifying important resources as well as for assessing re-access effort it is important to not just consider a resource and the action on this resource in isolation, but to understand the interaction with related resources.

In the prototype that is proposed and developed within this deliverable, we cast the problem of computing memory buoyancy to the problem of computing the mental effort that is needed by a human to access a specific resource, thus stressing the close relationship with mental processes. How difficult the human find himself when trying to access an object will be used as a proxy for the memory buoyancy of the respective resource. For example, if a human actor accesses a photo every day, it will take him or her nearly no time to recall the location of the photo in his computer. Our approach is the continuation of the previous work described in the deliverable D3.1, which is based on the decay models of human remembering and forgetting. In line with existing approaches, such models must take into account the usage activities of information resources in the past, and devise a salient way to predict the accessibility probability of the item in the future [10, 31]. The dynamic assessment of the resource access model raises the following challenges:

1. How much does the way a resource is accessed affect processes in human memory? For example, if a resource has been interacted with for a period of time very recently, does it maintain the accessibility in the human actor's memory better than other resources that got more frequent interactions, but less recently? What role does the document type play in the way human access the resource, and thus its memory buoyancy (for instance, do humans tend to locate old photos quicker than a text document)?

2. Can contextual information be helpful in complementing the assessment of the resource accessibility and importance? It is intuitive that using resource access time only can lead to poor performance, for the cases where resource interactions are

not explicitly observed but links are still activated in human mental process (for instance, accessing a picture of Edinburgh might trigger the memory of a picture from another trip to Edinburgh).

3. How to build a standard testing data for the evaluation of different methods and systems? The construction of the such a collection must be agnostic to any model or algorithms proposed. Some related work includes the analysis of different factors in email re-finding, which deems to be applicable to other domains as well [16]. For one part of this work, as we focus on the semantic desktop domain, where items represented in RDF models with relationships between their attributes, it is desirable to update the baseline models with these new data and investigate how they improve the overall system performance.

4. What is a good way to present the human memory buoyancy of known items so as to enable users to verify the system performance effectively and intuitively? Although this is not a core part of the research work, an user-friendly visualization of the memory buoyancy can greatly help get insights into the several aspects of the model, how it works, what is still missing, and which information can still be used for its improvement?

2.2.1 Resource Re-access Model

In order to tackle the above questions, we started with a simplified accessing model that takes into account the access time of resources, their context as encoded in the relationships between them and other entities, as well as implicitly captured in the activities of other resources. The general model is described informally as follows.

1. The actor starts accessing the information management system with an information processing need. This need involves an existing resource (e.g. reviewing the file content, copying or re-organizing the file to other location).

2. The actor first tries to recall the resource in isolation from other resources, infer the place of the resource from its properties only.

3. The actor can associate the resource of interest to other event or topics that are related to the resource (e.g. the holiday). The actor then retrieves one resources of the same event / topic by some cues (for instance, the folder named after the holiday in the computer, or files with creation time lying within the holiday time span)

4. If the resource is inter-connected to other resources, the actor can use the associative memory to trace along these resources as well. For each such resource, the actor recalls his or her most recent access activity.

We further hypothesize that the two last steps in the model stated above are picked up arbitrarily based on a pure heuristic system, instead of via a deep reasoning process. This is in line with the Kahneman's famous findings of fast thinking mechanism [25], where a heuristic, undeterministic memory system is always triggered before an educated, rational system is to be evoked. To this end, we propose a simplified model in which the human actor randomly chooses a strategy to re-access the resource under certain probability, and follow the strategy to locate the physical data objects. In the following, we describe the different components in our model that together form a fully generative process in resources re-accessing evaluation algorithm.

Time-Decay Model

In the previous deliverable, we have proposed a monotonic model based on Ebbinghaus forgetting function families. The choice of each forgetting function will be driven by how well it performs when applying to digital object remembering scenarios, as well as its complexity level (how difficult the model can be learned in practice). Too simplified model can easily learned and applied, but might fail to cover all salient aspects in today's human modern working and lifestyle environment (with the continuous support from computer systems), and vice versa.

Case study-Weinbull model. Here we revisit the model and refine it under the scenario of accessing resources in PIMO system. For the first implementation choice, our abstract model in D3.1 is revised based on recent findings in the performance of different forgetting functions in cognitive-based information retrieval [37], in which among other forgetting functions, Weinbull distribution-based function is suggested to outperform other priors in the area of retrieving information. We adapt our model to the extended Weinbull distribution as follows. For a given resource of interest r, let $d(r)$ denote all the data objects representing r. For each $d_i \in d(r)$, let t_i denote the timestamps of last access to d_i by the human actor. The accessibility of the resource r with respect to given timestamps t is defined by:

$$MB_{\mathbf{T}}(r,t) = b + (1-b)\mu e^{\sum_i - \frac{a\delta(d_i,t)^s}{s}} \tag{2.1}$$

parameter a measures the overall memory capacity of the system (how many data objects of the same type with d_i's that the system can store). Parameter s is one parameter of the Weinbull distribution and indicates the steepness of the forgetting function, i.e. how easily the system loses track of its member data objects. Parameter μ estimates the likelihood of initially storing the image of the resource in the short and long term human memory, and b is a asymptotic parameter with respect to the resource r (i.e. how inherently memorable the resource r is at $t = 0$?). The subscript $\mathbf{T}$ indicates that the memory buoyancy is calculated from time-decay model only.

In order to learn the model, it is noteworthy that b and μ are determined by non-temporal features related to the resources or human actor (e.g. resource type or properties, authorship of the resources and the human actor, etc.), while a and s depend on the characteristics of the information system itself. Given a training data set with such sufficient feature inputs for these two types of parameters, one can easily estimate their posterior values using a simple gradient-based procedure. The setting in which training data set is built and the interface for estimation and evaluation of the system is currently under the development, and will be reported in subsequent deliverables.

Propagation Model

While the time-decay model simulates the re-access model when a resource is seen in isolation (see above), it is more intuitive that in practice, humans tend to associate the resource of interest by other contextual information and attempt to recall the demanded data objects through whatever resource / entities that is easiest to reach (i.e. with highest accessibility value). To this extent, we propose a propagation model, inspired by the Successor model [43] and the PageRank algorithm [36] in the following way. Given a resource of interest r, we can go back to other entities related to r via some specific relations. Here not all relations are considered, some relations will make no impact on resource's accessibility. For example, authorship relation is agnostic to the propagation - access to the profile of an actor will not enable humans to remember his created files better. Therefore, we need a set of rules to define which relations will be included in the propagation model. For each such relation p, we can establish a number of links to r from entities e_i. Each entity e_i can contribute along p to the increase of the accessibility and, thus, the memory buoyancy of r. For example, if e_i is resource associated with a folder A, r is associated with one file in A (p is the "contains" relation), then each time an actor accesses to A, he can revive his old memory about the location of the file of r. However, if A contains many other files, this contribution will be suppressed by the *interference effect*[3], the actor's memory towards r will be mixed with other resources. In other words, the propagation of memory buoyancy of e_i (with respect to p) to that of r will be inversely proportional to the number of entities reachable from e_i via the relation p. We formulate this as follows:

$$MB_p(r,t) = \frac{1}{|p(?,r)|} \sum_i \frac{MB_p(e_i,t)}{|p(e_i,?)|} \tag{2.2}$$

where $|p(?,r)|$ and $|p(e_i,?)|$ indicate the number of entities e_i that are linked to r, and the number of entities reachable from e_i respectively (i.e. the number of RDF facts with the predicate equal to p, and the object or subject equal to r or e_i). The factor $|p(?,r)|$ is what differs our model from the traditional PageRank computation. In fact, it is not only for normalization over the sum, but also based on an interesting idea about human associative memory: People do not explore all the clues to locate a resource, they just pick up one arbitrarily and follow the traits. In equation 2.2, the accessibility to the resource r can be hinted via one of the entity e_i, each with a likelihood $\frac{1}{|p(?,r)|}$.

To propagate over multiple relations, we just have to notice that a human first chooses arbitrarily a strategy (e.g. traversing directory structures, opening dedicated applications,

[3] http://en.wikipedia.org/wiki/Interference_theory

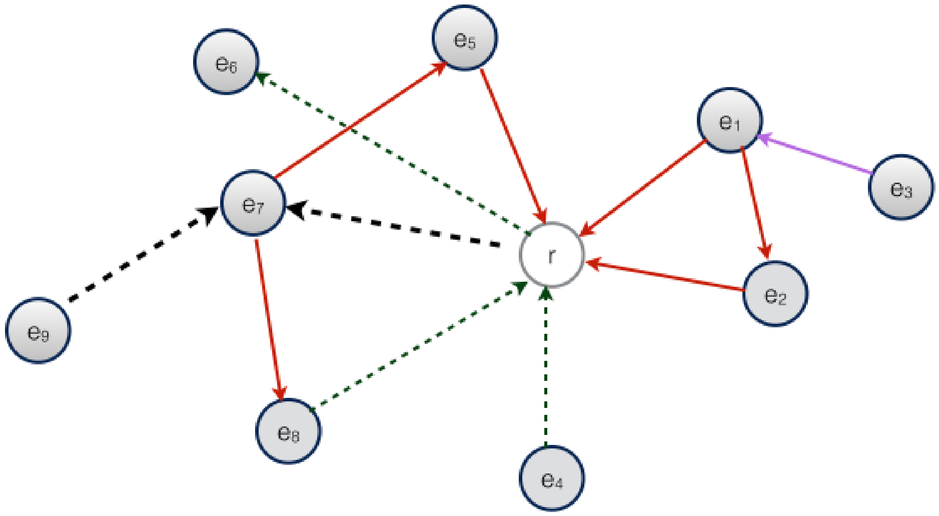

Figure 4: Illustration of the propagation of Memory Buoyancy along the heterogeneous resource relation graph

etc.), each with some prior likelihood. He then follows the corresponding hints to locate the resource. Thus we have the following measure of memory buoyancy as propagated from all other relations and resources:

$$MB_{\mathbf{P}}(r,t) = \sum_{p} \alpha_p MB_p(r,t) \tag{2.3}$$

where α_p indicates the weight of the relation p if it is used to propagate memory buoyancy from other resources to r. Solving the equation (2.3) means learning the weighting of α's from training data over different users and groups, constructing a graph of multiple relations among the resources and entities in the information space, and iteratively calculating the memory buoyancy until the values reach a stationary state. In practice, to avoid the long computation time, we can stop after a fixed number (e.g. 1000) of iterations.

Example As an illustrative example, in Figure 4 we compute the memory buoyancy of the resource r as propagated from other entities via different relations (each relation is represented by one style of edges). We start first by using the time-decay model to calculate the isolated memory buoyancy values for all nodes. Then, along the red edges, r receives half of memory buoyancy value from e_1 and full values from e_2, e_5, each with the likelihood $\frac{1}{3}$. Along the green dash edges, r receives full memory buoyancy values from e_4, e_8 with likelihood $\frac{1}{2}$. At the same time, along the bold dash edges, r propagates its previous value to e_7, etc. The computation repeats with the updated values of all nodes, until the graph reaches its stationary state (all values do not change), or after 1000 iterations.

2.2.2 Building Relation Graph from the PIMO Ontology and Interaction Log

In this deliverable, we focus on the personal scenario, and work primarily with the PIMO semantic desktop systems. For our model, it is crucial to build a relation graph between PIMO resources based on their relations. Two main sources of input are exploited here:

- The ontology of personal resources (how resources are linked together, properties of the resources, human-annotated info of the resources such as Events or Topics).
- The interaction logs of the resources.

In addition, we introduce three types of relations that can be included in the propagation model as follows:

1. **Resource - Container Relation** For example, whether a PIMO resource are subsumed in other resource by a containment relationship (e.g. documents in folder / Topic / Collection / Event / Job)
2. **Resource - Resource Relation** Capture the explicit relationships between resources via their RDF attributes (for example, a Reply-Email and an Original Email)
3. **Resource Temporal Relation** Inspired by the Successor model [43], we aim to see the resources that are accessed frequently within a time window as to belong to one short-term task, since for personal context, accessing multiple files in the information space within a short time period implicitly indicates that the user is focused on a certain task, and thus the files can be connected by an indirect tie w.r.t to such task. The remaining challenge is how we can formulate such temporal correlations in the domain of semantic desktop.

2.3 Prototype Components for Information Assessment

2.3.1 Prototype Overview

In Figure 5 we present a work flow of the forgetting process. The first component is the scheduler which is starting the process. Further, the metadata and the user interaction logs are loaded from a global repository, called metadata repository. For the computation of the memory buoyancy value, we also load the statistics and historic values. Finally, the results are stored in both repositories. In the following section, we will focus on the process inside the memory buoyancy assessor, which is part of the Forgettor component.

We describe the prototype of the computational framework for our memory buoyancy assessor. The framework here details the Forgettor components and focuses on the user information space (PIMO) with semantic data format. The assessment process can be triggered either by the active system (e.g. in reaction to a user request) or by the analyzer scheduler (which is not part of the Forgettor, but of the Scheduler component of the

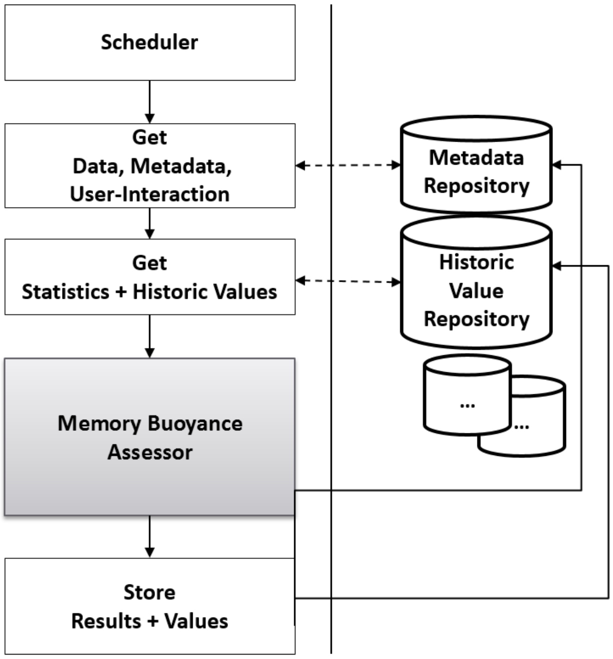

Figure 5: Workflow of the Forgetting Process

PoF middlelayer). The Propagation Memory Buoyancy Assessor, the main component responsible for calculating the memory buoyancy values of resources within the information space based on the propagation model, queries over the strategy repository to get the corresponding access methods to the information space. In addition to forgetting strategies, this repository stores information such as the URI address of the services by which the user data are sent to the Forgettor, or the authentication data for the Forgettor to successfully connect to the information client.

The client sends different types of data to the Forgettor. The first type -“low-dynamic data”- is the ontology, which represents the concepts and semantic structures of the information space (policy, rules, actor meta-data, meta-data about the systems in general) of the PIMO clients. The second type -a “medium-dynamic data”- consists of PIMO instances representing resources as well as their meta-data such as relationships between different resources. This two types can be sent periodically or on demand of the Forgettor. The third type -“highly-dynamic data”- captures the activities of the human actor on the data objects representing the resources. The log is sent regularly in batches to the Forgettor. The Forgettor then constructs the resource relation graph based on the three types of data received. Here not all relations are extracted and fed into the graph - the rule repository controls which relations will be considered or skipped (see Section 2.2.1). The Forgettor uses the inputs from the highly-dynamic data to calculate the isolated memory buoyancy values of the resources using the time-decay model; the calculation is triggered either by the analyzer scheduler or by the propagation MB assessor to initialize its loops. Finally, calculated memory buoyancy values are stored in the historic value repository, which can

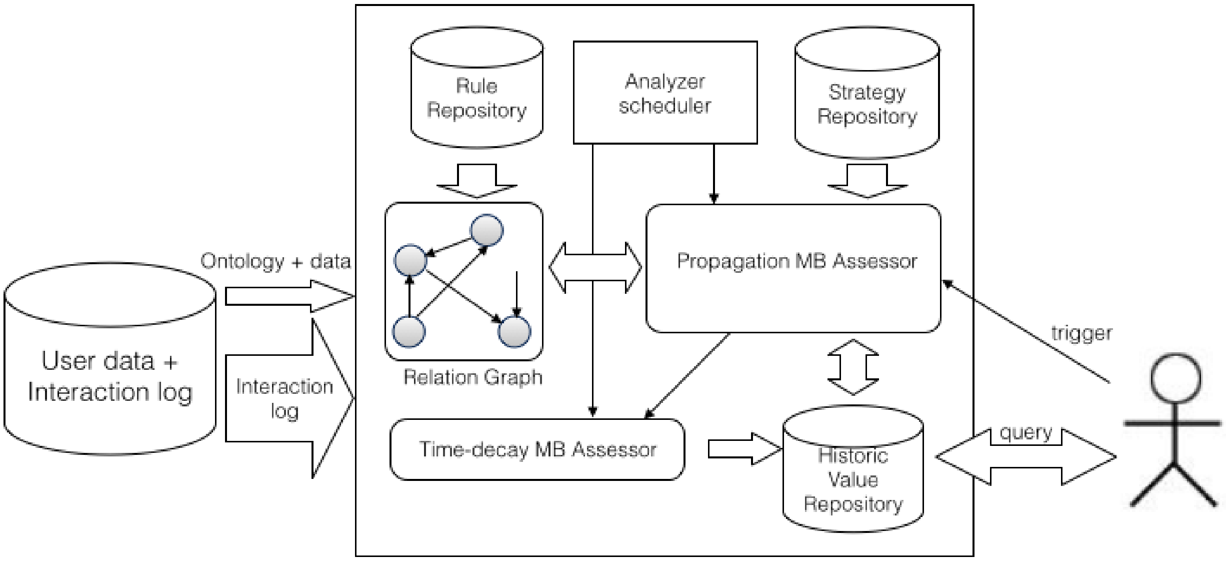

Figure 6: Computing components of Memory Buoyancy Assessor

be queried later by the active system via respective RESTful web services. The historic value repository is also used to materialize the time values calculated by the time-decay MB assessor, which can be queried by the propagation MB assessor later to improve the efficiency.

The implementation of the above framework has been started with the communicating RESTful web services between the and the PIMO system, and is now continued with the analyzer and the assessors. More detailed results will come in the subsequent deliverable. We are also developing different methodologies for evaluating the framework, based on both human studies and on the quantitative metrics.

2.3.2 RESTful Web Service

Interaction Log This web service are developed to enable the client (PIMO or TYPO3 system) to send the interaction logs periodically to the forgettor. The client calls the service each time it pushes the data to the cache of the forgettor, with format described below (Table 2), and it receives a response code acknowledge the status of the data importing. The data are sent in batch, each batch consists of several log entries for the performance sake. Each log entry conforms to the schema of the PIMOAction, as described in Section 2.1.3 (Interactions).

URL	http://forgetit.l3s.uni-hannover.de:99801/cache/pimo/log
Web service type	PUT
Parameter	None
Output	(1) OK - Data successfully imported; (2) SERVER_ERROR - the internal server error occurs; (3) FORMAT_ERROR - the importing is failed due to the invalid input format

Example input	`{"timestamp":1336373204762,"application": "MS-PowerPoint","userUri":"pimo:1327593979868:1", "conceptUri":"pimo:1332497855250:7","action": "access","actionObjectType":"File","occurrenceUri": "file:C:/Users/Heiko Maus/Desktop/ForgetIT-DFKI Vision.pptx"}`
	`{"timestamp":1336373290800,"application": "MS-PowerPoint","userUri":"pimo:1327593979868:1", "conceptUri":"pimo:1332497855250:7","action": "close","actionObjectType":"File","occurrenceUri": "file:C:/Users/Heiko Maus/Desktop/ForgetIT-DFKI Vision.pptx"}`
	`{"timestamp":1363361592732,"application":"Pimo", "userUri":"pimo:1363168140727:1","conceptUri": "pimo:1363168140727:8","action":"add", "actionObjectType":"Type","occurrenceUri": "http://www.dfki.de/web/living-labs-en/living-lab-virtual-office-laboratory"}`

Table 2: Format of the Interaction Log Web service

3 Information Value Assessment: Case Studies

A variety of features can be considered for the computation of memory buoyancy and preservation value. For our initial models of information value assessment in managed forgetting (Section 2) to work effectively in practice, it is crucial to study the impact of different features onto the effective learning in both memory buoyancy and preservation value models. This follows up with several research questions, including:

1. How can we effectively incorporate temporal features with non-temporal ones?
2. What features can best reflect the social aspects of human remembering and forgetting?
3. Given a (possibly) large amount of features, and given the time constraints for different components (MB assessor, querying components, etc.), what can we do to speed up the learning phase, for instance focusing on the most important features for some particular task?

In this section and the next section, we report on our case studies in an attempt to answer some of the above questions. Section 3.1 discusses our insights in the incorporation of temporal features into traditional feature-based settings, here in information retrieval scenarios. Section 3.2 reports our first analysis of the impact of features from the Social Web domain (such as social interaction) on information value assessment for managed forgetting. Section 3.3 reports some of our initial research in how to efficiently handle with large number of features, which will probably be relevant for both MB and PV assessment learning models.

3.1 Effectiveness of Temporal Features

3.1.1 Motivation

In order to assess the effectiveness of temporal features, we report this study on temporal features in the context of search result diversification. This is a common technique for tackling the problem of ambiguous and multi-faceted queries by maximizing query aspects or subtopics in a result list. In some special cases, subtopics associated to such queries can be temporally ambiguous, for instance, the query US Open is more likely to be targeting the tennis open in September, and the golf tournament in June. More precisely, users' search intent can be identified by the popularity of a subtopic with respect to the time where the query is issued. In this work, we address search result diversification for time-sensitive queries, where the temporal dynamics of query subtopics are explicitly determined and modeled into result diversification. Unlike aforementioned work [1, 7, 9, 14, 39, 41] that, in general, considered only static subtopics, we leverage dynamic subtopics for result diversification by analyzing two data sources (i.e., query logs and a document collection).

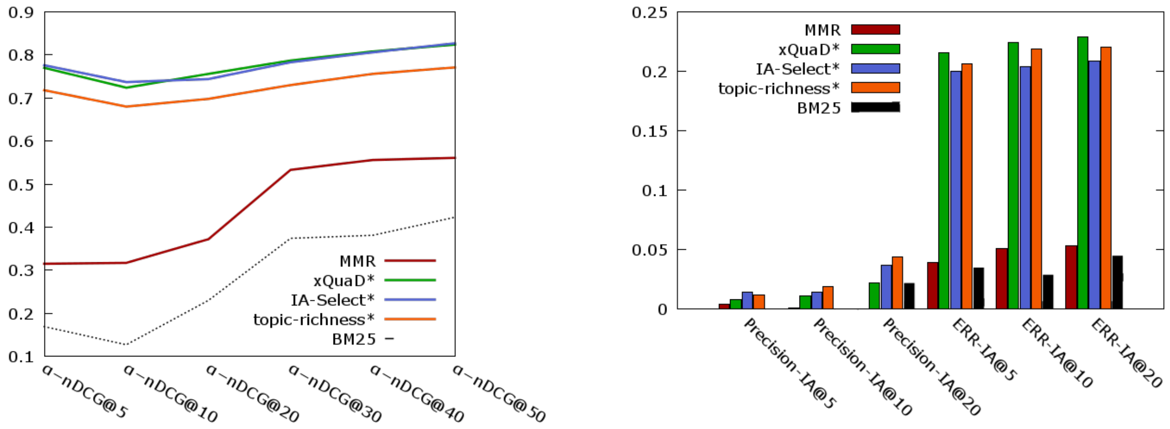

Figure 7: Ranking results of baseline models, * models are with dynamic subtopic mining

We analyze the temporal variability of query subtopics by applying subtopic mining techniques at different time periods. In addition, our analysis results reveal that the popularity of query aspects changes over time, which is possibly the influence of a real-world event. The analysis study is based on two data sources, namely, query logs and a temporal document collection, where time information is available. To this end, we propose three different time-aware search result diversification methods (namely, *temp-xQuaD*, *temp-IA-Select* and *temp-topic-richness*) which leverage dynamic subtopics and show the performance improvement over the existing non time-aware methods. The key idea of these methods is a recency and popularity-favor objective function of diversification. Readers can refer to [34] for more details.

3.1.2 Evaluation

State-of-the-art Model Performance We measure the performance of the following four state-of-the-art models: MMR, xQuaD, IA-Select and the topic richness model. The results are shown in Figures 7. For xQuaD, IA-Select and topic-richness, we use the mined temporal subtopics and their temporal weights as input (we skip their static methods (e.g., via Open Directory Project) since it is irrelevant in our case). We denote this change to the models with (*) symbol. We observe that xQuaD*, IA-Select* and topic-richness* outperform MMR (no account for subtopics), while MMR shows certain increase over the baseline where there is no diversity re-ranking.

Diversification Performance In these experiments, we aim to evaluate our time-aware models to answer our stated research question whether taking time into account that favors recency can improve the performance of the state-of-the-art diversification models. Tables 3 and 4 represent the results of the state-of-the-art and our time-aware models for α-nDCG and the two metrics Precision-IA and ERR-IA at different cutoffs respectively. Overall, our time-aware models exceed their original state-of-the-art diversification models in most of the experimental settings. temp-xQuaD is the most consistent algorithm that outperforms xQuaD and gives better results among the six tested algorithms. On the

Table 3: α-nDCG results with $^{\triangle}$ ($p < 0.05$), $^{\triangle\triangle}$ ($p < 0.01$) indicate a significant improvement

	$\alpha - nDCG@5$	$\alpha - nDCG@10$	$\alpha - nDCG@20$	$\alpha - nDCG@30$	$\alpha - nDCG@40$	$\alpha - nDCG@50$
temp-xQuaD	**0.783**$^{\triangle}$	**0.737**$^{\triangle}$	**0.758**$^{\triangle}$	**0.805**$^{\triangle\triangle}$	**0.820**$^{\triangle}$	**0.847**$^{\triangle}$
xQuaD*	0.699	0.687	0.706	0.751	0.772	0.789
temp-IA-Select	**0.781**	**0.739**$^{\triangle\triangle}$	**0.755**$^{\triangle\triangle}$	**0.798**$^{\triangle\triangle}$	**0.822**$^{\triangle\triangle}$	**0.836**$^{\triangle}$
IA-Select*	0.738	0.698	0.718	0.760	0.790	0.807
temp-topic-richness	**0.697**	**0.662**	**0.686**$^{\triangle}$	**0.731**$^{\triangle}$	**0.753**$^{\triangle}$	**0.769**$^{\triangle}$
topic-richness*	0.654	0.638	0.660	0.702	0.727	0.741

Table 4: Precision-IA and ERR-IA results with $^{\triangle}$ ($p < 0.05$) indicates a significant improvement

	P-IA@5	P-IA@10	P-IA@20	ERR-IA@5	ERR-IA@10	ERR-IA@20
temp-xQuaD	**0.010**	0.011	**0.029**	**0.214**	**0.218**	**0.232**$^{\triangle}$
xQuaD*	0.008	0.011	0.021	0.206	0.214	0.219
temp-IA-Select	0.010	0.010	0.027	**0.207**	**0.216**	**0.235**
IA-Select*	**0.013**	**0.013**	**0.034**	0.014	0.194	0.198
temp-topic-richness	0.010	0.011	0.030	**0.191**	**0.196**	**0.201**
topic-richness*	**0.011**	**0.017**	**0.040**	0.181	0.188	0.193

other hand, even though surpassing the base model, temp-topic-richness gives a lower performance compared to the other two time-aware diversification models. However, the model is meant for taking subtopics from multiple sources. Its performance could be enhanced if we account for other sources of subtopics (i.e., query log).

3.2 Effectiveness of Social Features

3.2.1 Motivation

In the previous deliverable D3.1, we have motivated a temporal summarization for Social Web posts. The summary includes activities, events, interactions and thoughts of the last months or years. It can also be used for personal reminiscence as well as for keeping track with developments in the lives of not-so-close friends. One of the core challenges of automatically creating such summary is to decide which posts to *remember*, i.e., consider for inclusion into a summary and which to *forget*. Keeping everything would contradict the idea of a summary and would also neglect the often intentionally ephemeral nature of Social Web posts.

As a first step for this selection process, we extract and analyze the most impacting features that characterize memorable posts [33]. Our experimental work is based on a user evaluation for discovering human expectations towards content retention. The goal of our analysis is to identify core features which can be used to classify memorable posts with high effectiveness. We also show that the identified feature set outperforms the usage of core social features alone.

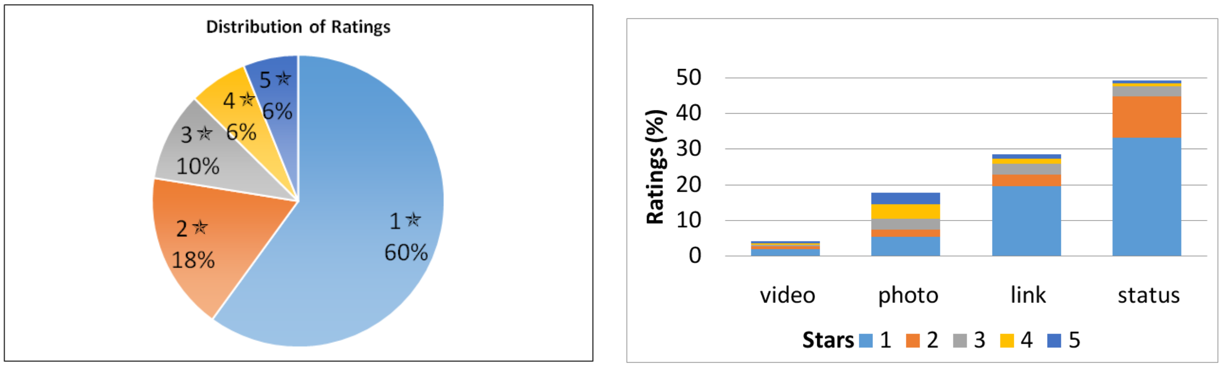

Figure 8: Distribution of ratings of posts.

3.2.2 Evaluation

In order to build a ground truth of memorable social network posts, we set up a user study on top of the Facebook platform. The goal of this evaluation was to collect participants' opinions regarding the retention preferences for their own Facebook posts, shares and updates. In this subsection we shortly describe the dataset collected from the user evaluation and an overview analysis of the data. In total, we had 20 participants, 15 male and 5 females ranging from age 25 to 37. Together they evaluated 3,330 posts. Additionally, once the user provided us authorization to access their data, we were able to collect general numbers that helps us to depict the general use of Facebook Social network.

In our user study, each participant had to judge their own posts on a 5-point Likert scale answering the following question: *How relevant is your post for future reference?* We asked participants to judge at least 100 of their posts. It is important to note that we are not judging participant's memory skills. Instead, we are collecting their personal opinion. Due to that, we presented the participants' posts in a chronological order starting from the most recent. For active users, 100 posts may date back to just a few days, reaching up to months for the less active ones.

Figure 8 depicts the distribution of the ratings. We clearly identify the dominance (60%) of irrelevant posts (1 star). Further, Facebook defines seven types of posts, namely: link, checkin, offer, photo, question, swf and video. This basically describes the type of content that is attached to a post. Figure 8 shows the distribution of posts among these categories. 49% of the evaluated posts consist of status updates, followed by shared links (28%), photos (17%) and videos(4%).

Figure 8 (right) displays the average ratings over time (from Jan. 2009 until Nov. 2013). This shows a clear trend where participants in the evaluation assigned higher ratings to more recent posts. This is in line with the idea of a decay function underlying the content retention model. From this statistics, one can deduce first ideas for the features that have a higher impact in the detection of memorable posts. Roughly speaking, recent photos with high number of likes and high number comments, seems to be the best evidence.

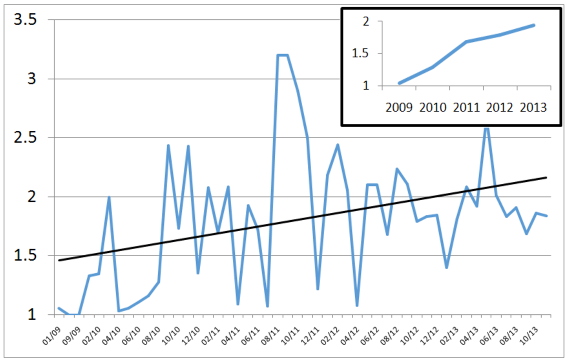

Figure 9: Average ratings over time grouped by month and by year (top right).

3.3 Efficiency of Information Value Assessment

3.3.1 Motivation

The number of features in search engine and social web can increase to several hundreds, it is desirable to identify a subset of features that yield a comparable effectiveness, in terms of efficiency and performance, to using all the features. In this study [32], we adopt various greedy result diversification strategies to the problem of feature selection for learning to rank. Our experimental evaluations using several standard datasets reveal that such diversification methods are quite effective in identifying the feature subsets in comparison to the baselines from the literature. Our methods can be applied to any other ranking problem with a known feature-impact-factor.

In a recent study, Geng et al. proposed a filtering-based feature selection method that aims to select a subset of features that are both effective and dissimilar to each other [19]. Inspired from this study, we draw an analogy between the feature selection and result diversification problems. In the literature, a rich set of greedy diversification methods are proposed to select both relevant and diverse top-k results for web search queries (e.g., see [8, 21, 47, 40, 42]). We apply three representative diversification methods, namely, Maximal Marginal Relevance (MMR) [8], MaxSum Dispersion (MSD) [21] and Modern Portfolio Theory (MPT) [47, 40] to the feature selection problem for LETOR.

To the best of our knowledge, none of these methods are employed in the context of learning to rank with the standard search engine datasets. In the next paragraph, we first describe the baseline strategies for the feature selection from the literature, and then discuss how we adopt the result diversification methods for this purpose.

Baseline Feature Selection Methods

Top-k Relevant (TopK): A straightforward method for feature selection is choosing the top-k features that individually yield the highest average relevance scores over the queries [13].

Greedy Search Algorithm (GAS): This is the greedy strategy proposed by Geng et al. in [19].

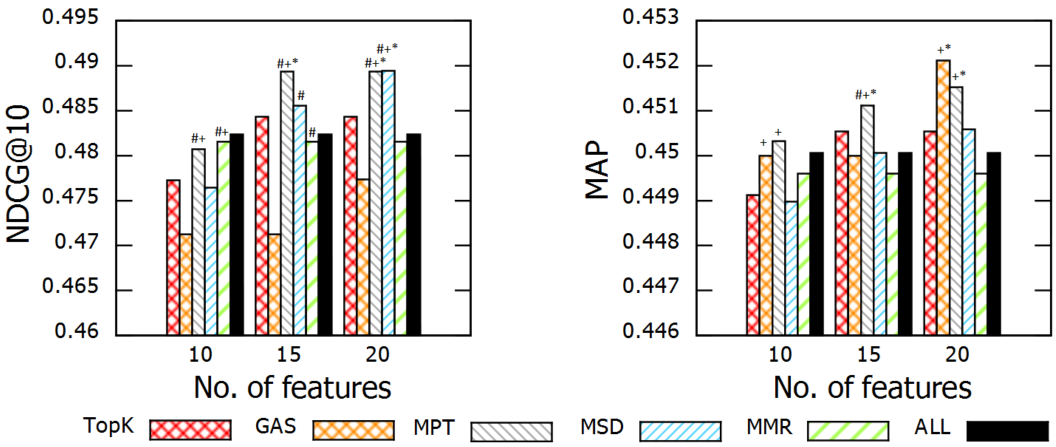

Figure 10: Ranking effectiveness on OHSUMED: NDCG@10 (left) and MAP (right).

Diversification Methods for Feature Selection

Maximal Marginal Relevance (MMR): This is a well-known greedy strategy originally proposed in [8]. In this study, we adopt a version of MMR described in [46].

MaxSum Dispersion (MSD): An alternative representation of the diversification (and hence, feature selection) problem is casting it to the facility dispersion problem in the operations research field [21].

Modern Portfolio Theory (MPT): This approach is based on the famous financial theory which states that one should diversify her portfolio by maximizing the expected return (i.e, mean) and minimizing the involved risk (i.e., variance). In case of the result diversification, this statement implies that we have to select the documents that maximize the relevance and have a low variance of relevance [47, 40].

3.3.2 Evaluation

Our experiments are conducted on the standard LETOR dataset, OHSUMED[4]. Our evaluations employ RankSVM [23], which is a very widely used pairwise LETOR algorithm. More specifically, we used SVMRank[5] library implementation. We trained the classifier with a linear kernel with $\epsilon = 0.001$. In Figure 10, we report the NDCG@10 and MAP scores obtained on the OHSUMED dataset using the baseline and proposed feature selection methods. We observe that when the number of selected features is greater than 10, the performance is comparable or better than using all features (ALL). Furthermore, the methods adapted from the diversity field outperform the baselines (TopK and GAS).

The statistical significance of our methods is verified using the paired t-test with $p < 0.05$. In Figures 1-3, we show the significant differences to the baselines TopK (denoted with +), GAS (denoted with #) and ALL (denoted with *).

[4] http://research.microsoft.com/en-us/um/beijing/

[5] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

4 Complementing Human Memory: Case Studies

One of the main research challenges in ForgetIT is to understand to which extent human can remember details or general shape of an event in their real life. The goal here is not to simulate this mental process, but to provide a salient way to complement such process. ForgetIT technologies approach this issue via various ways to assess the information value of digital objects, in which memory buoyancy component design and implementation matters have been proposed and discussed in Sections 2 and 3. However, it is still on demand how the Forgettor can base on these values to support a decision making regarding preserving resources in personal or organizational settings.

In this section, we continue our study on complementing human memory from the previous deliverable. We focus on the understanding of complementarity in episodic memory, among other types. Although the sources of such memory vary vastly from individual to individual, in the larger global scale, we believe that there are some common features that govern the human remembering towards public events. Getting insights into such features can greatly help computers to measure preservation value of digital objects, by associating it with different events. We focus our study on two case studies. First, we see how social media can be seen as the extended human mind when tracking and recalling recent public events. Second, we see how human memory towards past events can be triggered via different similar ongoing events.

4.1 How Social Media Complement Human Memory in Public Events: Case Study of Wikipedia and Twitter Trending Topics

4.1.1 Motivation

With the recent proliferation of a vast number of social media platform (Facebook, Twitter, Tumblr, Reddit, etc.), users now have a variety of choices to get informed and follow real-world events of interest. However, it is not clear how users' collective attention towards different types of events varies from source to source. In this case study, we make the first attempt to answer this question by mining two huge public event resources: Twitter[6] and Wikipedia. It is widely believed that while Twitter has been a great source of up-to-date information about real-world incidents, it is also contaminated by lots of spam topics, such as endogenous information that disseminate within Twitter community only, without obvious real-world incident matching. Information in Twitter often exhibits spikes during prominent events such as Super Bowl, therefore existing methods detect and track real-world events reported in Twitter typically through the volume of posts [45, 49]. However, the lack of contextual information from resources other than Twitter sphere makes these methods unable to identify whether trending topics truly reflect real-world events, or just a “virtual” topic such as “#uFromLAif” (which was a spontaneous memes staying within

[6] http://www.twitter.com

Twitter only). This makes systems misled with spam topics, while possibly missing other potential events In the meanwhile, Wikipedia has increasingly become a creditable source of knowledge about scientific as well as person-related and event-related information. Recent work suggests that trending in Wikipedia article views or edit activities can also be a signal of new real life events [20] (with the time lags estimated to a few hours [35]). The relevance and precision is intuitively higher as compared to ones in Twitter, as information in Wikipedia is more creditable and focused. In this work, we propose using Wikipedia to improve the analysis of trending topics in Twitter. As in previous work [45, 49], we propose to rely on hashtag to predict the future behaviour of trending topics in Twitter. Unlike previous work, we predict how long it takes for a trending topic to saturate after the peak. As shown in our experiments, saturation length is a stronger signal for indicating the long-term influence of a hashtag than just peak volumes, and better distinguishes endogenous from exogenous topics.

4.1.2 Methodology

Datasets For the Wikipedia data, we obtained the English revision history dump on 30 Nov. 2012 (380 million updates of 4 million articles), and the Wikipedia page view count statistics dataset. The Twitter dataset is TREC Tweets2011 corpus[7], which contains 16 million public tweets sampled from 23.01 to 08.02.2011.

Burst and Saturation To define a trending hashtag, we employed the simplified *Rapid Rising* strategy [18]. For each time point t with the value $n(t)$, we look back at preceding k values, and claim t a peak if the current value is l-time standard deviations higher than the mean value of the preceding window: $n(t) \geq l\sqrt{(n(t-i)-\mu)^2} + \mu, i = \overline{1,k}$ where μ is the mean of k variables $n(t-i)$. We measure the saturation length as the number of days from the first peak to the closest day where the hashtag volume goes under a threshold τ. If the hashtag has several peaks, saturation length is the average duration. We observe that $k = 3, l = 3$ and $\tau = 10$ give the most intuitive peak outcomes in Tweets2011.

We get only hashtags with more than 40 tweets in at least one day, and choose 628 random hashtags, amounting for 672,580 tweets. For each hahstag, assessors are displayed with the set of peak days and top 50 tweets on each day. The assessors then use keywords, mentions, abbreviations, etc. in the tweets and use the published days to issue to a search engine and Wikipedia. Each hashtag is annotated as whether the related information can be found on the Web (exogenous), and further whether it is found on Wikipedia (ongoing, otherwise breaking event). In the end, we have 275 hashtags about endogenous topics, 353 about exogenous topics, in which 231 are breaking events (information found on the Web but not in Wikipedia in the peak day) and 122 ongoing topics. Figure 11 shows the distributions of hashtag saturation lengths in endogenous and exogenous topics. The power-like curves agree with previous findings [3] that most Twitter hashtags decay very fast. Moreover, exogenous topics saturate longer than endogenous ones, with 10% saturating longer than 3 days compared to 2% in the endogenous set.

[7] http://trec.nist.gov/data/tweets

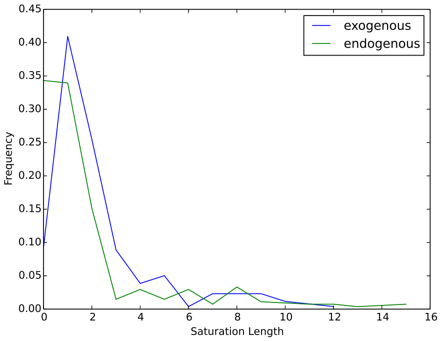

Figure 11: Saturation distributions in Tweets2011

Saturation Prediction We propose a framework that given a hashtag h peaked on day t_0, can predict the saturation length $L(t, h)$. As finding an exact value of L is difficult and often not necessary, we propose to classify the range which L falls in. For the Tweets2011 dataset (spanning 3 weeks), the range is defined as: [*1*] (last only 1 day), [*2-3*], [*4-7*] (last longer than 3 days to 1 week), [*7-14*] (last longer than 1 week to 2 weeks), [*14-21*] (last longer than 2 but less than 3 weeks), [*0 or 22+*] (last more than 3 weeks or no burst).

Entity Linking For each hashtag h and the peak day t_0, we concatenate all the tweets in the order of published time, and use existing tools to link to a set of Wikipedia entities. For supervised approach, we use WikipediaMiner [30], and for the unsupervised approach, we use TwiNER [28] to identify entities in tweets, and AIDA[8] to disambiguate the entities.

Model Features We define 40 features, grouped in four categories as described in Table 5. The hashtag and tweets types are derived from previous work [45, 49] and used as the baseline. We propose several features extracted from matching Wikipedia entities to enhance the contextual knowledge. For instance, the authority score of an entity measures how importance it is w.r.t. to other entities: $authority(w) = \frac{|IN(w)|}{|OUT(w)|}$, with IN and OUT are incoming and outgoing link sets of the snapshot of article w on day t.

Type	Features
Hashtag	(1) Hashtag length, (2) No. of segmented words in the hashtag, (3) (binary) if it has digits, (4) if it collocate with other hashtags, (5) no. of collocating hashtags, (6) fraction of capitalized characters in the hashtag
Tweets	(1)-(4) fraction of tweets having URLs/hashtags/ mentions/emoticons, (5)-(8) fraction of URLs/hashtags/ mentions/emoticons over tokens, (9) no. of distinct users, (10) average token length per tweet, (11) fraction of retweets, (12) 3-d emoticon vectors of tweets
Wiki static	(1) no. of matching Wikipedia articles, (2) no. of persons, (3) no. of locations, (4)-(5) maximal/average authority score of Wikipedia pages
Wiki Temporal	(1)-(4) if the edit/view count increase in all/any Wikipedia articles that match the hashtags, (5)-(8) minimum/maximal length of increase chains in view/edit count, (9) fraction of Wikipedia revisions that have URLs

Table 5: Features used for prediction

[8] https://github.com/yago-naga/aida

4.1.3 Preliminary Results and Discussion

Result Table 6 summarizes the accuracy of the classification for different feature settings. In the FullSet, for both entity linking systems, we see a clear improvement when incorporating Wikipedia information as features. Wikipedia edit history and Wikipedia structure information contribute the most to the increase in accuracy. Moreover, the performance of TwiNER+AIDA system is lower. This is explained by the fact that TwiNER is unsupervised and has inferior quality, and that AIDA is backed by the YAGO knowledge base, which only contains a subset of Wikipedia articles. Again, this emphasizes the importance of adding more information from Wikipedia to improve the prediction.

The performance varies in different kinds of trending topics. For endogenous topics, the result is unstable with both entity linking outcomes; adding different Wikipedia features sometimes harm the performance (although it does improve in general). This is because endogenous hashtags merely diffuse information within Twitter communities, and mentioned entities in tweets will thus not correlate well with the main content of the Twitter topic. For breaking topics, both systems do not gain any improvements with Wikipedia features, this confirms the fact that breaking events in Twitter spread quicker than in Wikipedia. For ongoing topics, incorporating Wikipedia information does effectively improve the performance of the prediction in both entity linking settings. Method based on WikipediaMiner performs best with Wikipedia static and edit features, and method based on TwiNER+AIDA performs best on the full combination. Finally, the general prediction performance of the systems can gain significant benefit when we increase the size of our data (from sample sets to FullSet). This positively supports the idea that despite the small size of the annotated dataset, our system does not overfit and has a good general ability.

	FullSet		Endo		Breaking		Ongoing	
	WM	TwiNER+AIDA	WM	TwiNER+AIDA	WM	TwiNER+AIDA	WM	TwiNER+AIDA
Baseline	0.5865	0.5865	0.4167	0.4167	**0.6333**	**0.6333**	0.5444	0.5444
wstatic	0.7242	0.6912	0.6234	0.6190	0.6292	0.5801	0.5667	0.5590
wview	0.7284	0.6976	**0.7382**	0.7146	0.6333	0.5711	0.5667	0.5616
wedit	0.7383	0.6882	0.7355	0.7192	0.6333	0.5825	0.5731	0.5684
wstatic+wview	0.7355	0.7012	0.5612	0.6018	0.6250	0.5804	0.5625	0.5718
wstatic+wedit	**0.7411**	0.7134	0.6345	0.6129	0.6250	0.5727	**0.5778**	0.5645
wview+wedit	0.7346	0.7035	0.7337	**0.7682**	0.6333	0.5705	0.5670	0.5691
wstatic+wview+wedit	0.7374	**0.7276**	0.4333	0.4212	0.6250	0.5793	0.5767	**0.5792**

Table 6: Accuracy of hashtag staturation prediction in Tweets2011

4.2 Understanding Collective Memory in Wikipedia

4.2.1 Motivation

The way humans forget and remember is a fascinating area of research, both for individual and collective remembering: Aspects such as the constructiveness of memory are challenging our intuitive understanding; forgetting enables us to stay focused and to cope with the multitude of or daily experiences; and the way past memories are triggered by new experiences is sometimes surprising.

In our analysis, we investigate the triggering or reviving of memories of past events using revisiting pattern in English Wikipedia as indicators for what is collectively (actively) remembered and what is rather on the path of forgetting. In general, individual memories are subject to a forgetting process, which is driven by some form of the forgetting curve first proposed by Ebbinghaus [15], which leads to a decay function with fast loss of details especially in the early phase after an event. Various factors can, however, boost human memory of a event or person from one's past, such as similar events, anniversaries or even a scent. Such triggering of memories can also be observed for more global events on a cumulative level of communities as the sum of individual remembering re-enforced by information sharing and media coverage. The *2011 nuclear catastrophe in Fukoshima* did, for example, trigger memories of the *Chernobyl event* happened 25 years before raising the Wikipedia event page views from about 9,500 views per day in the first two months of 2011 to up to more than half a million views per day at the time of the Fukoshima disaster (around March 15).

In more detail we are interested in the catalysts for such re-viving of event memory. We investigate, which role the time passed, the type of event, and other factors play in reviving memory. Our work extends the work of [4], who examine collective memory based on its reflection in a newspaper collection, in two directions. Firstly, we analyze the long-term dynamics of collective remembering by looking how forgetting is interrupted by memory revival. This also supplements work on the early memory construction phase in creating Wikipedia articles [17] by looking into long-term temporal development. Secondly, we add an extra perspective by analysing what people actually look at (in Wikipedia), complementing the News coverage perspective of [4].

4.2.2 Forgetting and Memory Catalysts

Remembering and forgetting in the context of high-impact events, so called *flashbulb memories*, have been analysed in various studies [24, 11, 12] in cognitive psychology. According to a more recent definition [11], flashbulb memory is “memory about an emotionally impacting event of personal and national importance, which is consequential, socially shared and rehearsed by media”. It comprises an autobiographical part, which refers to remembering the personal context, in which one learned about the event and the memory about the event itself. Aspects that have been studied are the details that people still remember over different periods of time (e.g. 1 week, 11 and 35 months after the event in [24]), the confidence and consistency of their memories over time and the impact of media coverage. However, due to their qualitative nature, those studies are typically limited to a small number of events and a restricted number of users.

Social media analysis have been successfully used in different works for analysing collective attention and awareness[27]. Due to their dynamics, events typically play an important role in such analysis. The transition to analysing remembering of events as a

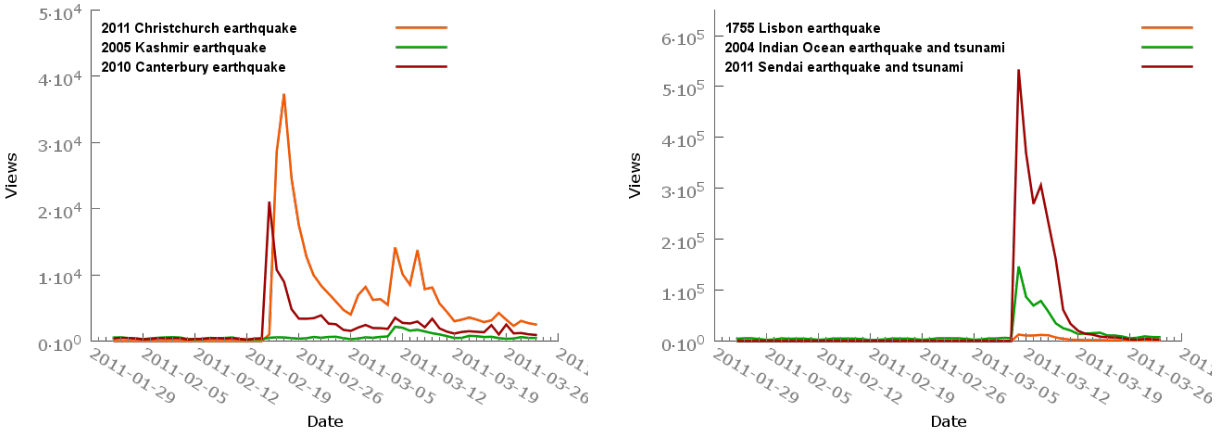

Figure 12: Views triggered by two events: 1) 2011 Christchurch Earthquake, and 2) 2011 Tohoku earthquake and tsunami.

crowd phenomenon relates individual remembering to collective remembering. In social sciences, the concept of collective memory [2, 22] is used in this context. It refers to the collectively constructed image (memory) of the past, which is shared by a community and, roughly speaking, is used by this community for framing their current understanding and activities as a community. The Web in general, and especially the Social Web has a high impact on collective remembering [38]. Within Social media, its popularity as an information reference and the easy and long-term accessibility of information about an event makes Wikipedia a promising subject for analysing collective remembering. In addition to the access numbers, the importance that is assigned to Wikipedia as an information reference for event information is confirmed by the high level of community involvement reflected in the number of editors (19 Mio registered users and about 30 thousand active editors[9] in English Wikipedia), the fast reflection of new events in Wikipedia [26], and the conflicts and *edit wars* that can be observed on controversial topics. Although religious and political topics are most dominant in edit wars, there is also a considerable number of events in the top 10 lists of controversial topics extracted from Wikipedia in different languages in [48].

Figure 12 (left) shows the views of the event page for the earthquake in Christchurch, New Zealand in February 2011 (as triggering event) and compares it with the view number of two other earthquakes namely the earthquake in Canterbury in September 2010 and the large earthquake in Kashmir in October 2005. The strong peak in the views of the Canterbury earthquake around February 20 suggests a strong influence of the Christchurch earthquake as a catalyst for remembering the Canterbury earthquake. This strong influence can be explained by the facts that a) both earthquakes happened in the same region and b) there is a time delay of just five months between the two events. In contrast, memory for the Kashmir earthquake, which is more distant in time and location, seems to be revived to a much lesser degree by the Christchurch earthquake.

Figure 12 (right) shows page views for the event page of the 2011 Tsunami in Japan

[9] editors with more than 5 edits per month

as the triggering event and views for the event page for the Indian Ocean Tsunami in 2004. Again the view numbers suggest that the event in Japan acts as a catalyst for remembering the 2004 Tsunami, and -taking a closer look to Figure 12 (left) - also to the event pages of both earthquakes in New Zealand. Interestingly, there is an increase even for earthquakes, which lay far more in the past as the Lisbon earthquake in 1755 shown as the third line in Figure 12. Of course, an increased number of Wikipedia views is only an indirect signal of memory revival for the considered event. However, we believe that a person, who visits an event page from a past event at least thinks of the event, which brings it back to active memory. Furthermore, visiting a Wikipedia page on the event on purpose will typically result also in reading some information about the event such refreshing or extending the information memorized about the event.

Given this first observation, our plan for the next step is to employ time series analysis: (1) temporal correlations in peaking page visits between events, (2) a surprise score or the residual sum of squares on prediction error, and (3) the skewness of view shapes, as indicators for the capability to act as a catalyst for the memories about the past event. Furthermore, we will investigate if there are also other indicators of relationships between the events (e.g. the same types or magnitude of events, same city or country, etc.), by using different features, namely, time, location and impact.

5 Research Plans and Future Activities

In this section, we outline our plans for the next steps in WP3 including research plans as well next steps towards realizing the concept of managed forgetting.The research plan and the plans for further implementation activities will be frequently revisited and re-aligned with the activities in the rest of the project as well as with the requirements identified in collaboration with the two application pilots (WP9 and WP10) and the work in the architecture work package (WP8).

5.1 Information Assessment in Photo Preservation Scenarios

5.1.1 Motivation

Initial idea for modeling managed forgetting have been presented in Section 2.2, which is focused on modeling memory buoyancy (MB) through a accessibility model. Insights into assessing preservation value (PV) is still under study, in order to complement with memory buoyancy assessment in a unified managed forgetting framework. In the next months we will work on progressing in this area, gathering further insights on human expectations, and defining methods for computing MB and PV, and how to combine them in more detail. To further investigate the contribution of different factors in driving human assessment of resource preservation value (personal and organizational settings), as well as to foster the design of PV assessor models, we plan to conduct systematic experiments with human explicit and implicit feedback on photo preservation scenario (see Table 1, deliverable D3.1). The idea is that we design a prototype that enables human to look at their photo collections stored in their computers, annotate and send the feedback to the system for the analysis in an anonymous way.

5.1.2 Methodology and Evaluation Plan

For this study, we use the personal photo collections of volunteers and build a dataset of event-related photo albums of different ages (i.e. albums that were taken several years ago, as well as taken recently). To cope with the privacy issues, we design our dataset in a way that it does not require contributors either to share their photos to others, and the photos are kept in original place of the authors' choice without additional copies, e.g., in centralized databases or corpora. This characteristic means that the dataset does not reveal any sensitive information from the contributors, and thus enables them to share more of their photo albums and experiences.

In short, the dataset consists of two parts: *reference part* and *content part*. The reference part is a set of references to the location of personal photo albums. For instance, the location can be an absolute path of the directory where the photos are stored. The reference will be encoded and synchronized between the client device (e.g., cell phones and computers) and the server, so as to guarantee both the anonymity of the collections as well as to keep information up-to-date. The content part is a set of the physical location of photo albums, as stored in the user's own device. It contains user photos, plus one extra file named manifest.txt in the root of the location. The manifest file consists of human-annotated data, and is generated through a user interface runs in the contributor's local machines. The interface is designated to enable humans to quickly annotate their photo albums with some properties such as: age (when the albums were taken), location (where the albums were taken), privacy levels (private, or can shared with close friends, or can be exposed to acquaintances, or can be public), people tagged in the photos, keywords describing the concepts represent in the photos (household, picnic, etc.). To ensure the consistencies of tagging among annotators, we will utilize a standard taxonomy such as WordNet [29], as adopted in existing image annotated dataset such as ImageNet [10]

Once the dataset have been built, a different models will be built to learn the way humans justify the preservation values of their photos, taking into consideration the features and training labels that were obtained from the manifest files sent from annotators' local devices. Should the models require content features (e.g. photo visual features such as SIFT, histograms, etc.), a separate computing component will be deployed and run in annotator's local machine, and only extracted features (mostly in numerical formats) will be sent to the centralized learning system.

5.2 Temporal Summarization of Social Web Content

5.2.1 Motivation

The creation, handling, and sharing of electronic information within the personal sphere has seen a unprecedented growth and change in recent years. Cornerstones for such development are new technical devices and corresponding changes in our everyday behaviors. In social networks like Facebook, Google+, and Twitter people share lots of different content about their personal life, interests and activities that are considered a valuable part of personal remembrance. The most of this shared information get a very short attention from the community and it gets forgotten. For a user it is almost impossible to get an overview about his activities and personal highlights over a long term in the past. Our idea is to provide a personalized summary from the social web content of the user from a particular time period. But the summary can also be helpful for other users to get a general overview of their contact's activity, e.g. by sharing the summary with family and close friends.

[10] http://www.image-net.org/explore

5.2.2 Methodology and Evaluation Plan

For our study we use the Facebook profile of a user with all his different types of content like photos, videos, comments, likes and relations for creating a summary of a particular time period (e.g. monthly or yearly). One challenge is to select a representative subset of the different types of content and consider their connections and relevance in a context. This differs from the most of the summarization techniques in social web which use only one type of information like text, tweets, comments or tags. We finalize our research question as follows: 1) *What are important types of content in the summary?*, 2) *How to identify highly relevant content from the personal perspective of the user?*, and 3) *How to visualize the summarized results?*

The first step of our evaluation is based on a questionnaire of Facebook users. We will ask general questions about user's Facebook activity and the content that should be captured in their summary. In the second part, users will have to evaluate their personal Facebook summary as well as the visualization of the summary.

5.3 Managed Forgetting Strategies

A further important activity in the coming months is to investigate, how to complement the developed information value assessment approaches for memory buoyancy and preservation value towards a full managed forgetting solution. This especially includes the definition and implementation of forgetting options and forgetting strategies.

6 Conclusions

This deliverable describes our proposed data models and framework for information value assessment. We defined the environment and actions in which forgetting and preservation scenarios and functionality are studied and highlighted. Based on concepts in artificial intelligence and Semantic Web, we represented *Resources*, *Interactions* and *Human actors* in an information space using ontology defined in the PIMO semantic desktop system. In order to support our proposed ideas, we conducted several studies in more general settings for evaluating the impacts of potential features for both memory buoyancy and preservation value. Specifically, we performed the evaluation of information value assessment in term of effectiveness and efficiency. In addition to address the aspect of complementing human memory, we reported our preliminary research results on complementing human memory. To this end, we outlined planned research activities towards realizing the concept of managed forgetting in the coming months of the project.

References

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of WSDM '09*, 2009.

[2] J. Assmann and J. Czaplicka. Collective memory and cultural identity. *New German Critique*, (65):pp. 125–133, 1995.

[3] S. Asur, B. A. Huberman, G. Szabo, and C. Wang. Trends in Social Media: Persistence and decay. In *ICWSM*, 2011.

[4] C.-m. Au Yeung and A. Jatowt. Studying how the past is remembered: towards computational history through large scale text mining. In *Proceedings of CIKM '11*, 2011.

[5] T. Berners-Lee, R. Fielding, and L. Masinter. Uniform resource identifiers (uri): Generic syntax, 1998.

[6] D. Brickley and R. V. Guha, editors. *RDF Vocabulary Description Language 1.0: RDF Schema*. W3C Recommendation. World Wide Web Consortium, Feb. 2004.

[7] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR '98*, 1998.

[8] J. G. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. of SIGIR'98*, 1998.

[9] B. Carterette and P. Chandar. Probabilistic models of ranking novel documents for faceted topic retrieval. In *Proceedings of CIKM '09*, 2009.

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

[11] E. Coluccia, C. Bianco, and M. A. Brandimonte. Autobiographical and event memories for surprising and unsurprising events. *Applied Cognitive Psychology*, 24(2):177–199, 2010.

[12] A. R. A. Conway, L. J. Skitka, J. A. Hemmerich, and T. C. Kershaw. Flashbulb memory for 11 september 2001. *Applied Cognitive Psychology*, 23(5):605–623, 2009.

[13] V. Dang and W. B. Croft. Feature selection for document ranking using best first search and coordinate ascent. In *Proc. of SIGIR'10 Workshop on Feature Generation and Selection for Information Retrieval*, 2010.

[14] Z. Dou, S. Hu, K. Chen, R. Song, and J.-R. Wen. Multi-dimensional search result diversification. In *Proceedings of WSDM '11*, 2011.

[15] H. Ebbinghaus. *Über das Gedächtnis. Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot, Leipzig, 1885.

[16] D. Elsweiler, M. Baillie, and I. Ruthven. Exploring memory in email refinding. *ACM Transactions on Information Systems (TOIS)*, 26(4):21, 2008.

[17] M. Ferron and P. Massa. Psychological processes underlying wikipedia representations of natural and manmade disasters. In *Proceedings of WikiSym '12*, 2012.

[18] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu. Parameter free bursty events detection in text streams. In *VLDB*, 2005.

[19] X. Geng, T.-Y. Liu, T. Qin, and H. Li. Feature selection for ranking. In *Proc. of SIGIR'07*, pages 407–414, 2007.

[20] M. Georgescu, N. Kanhabua, D. Krause, W. Nejdl, and S. Siersdorfer. Extracting event-related information from article updates in Wikipedia. In *ECIR*, 2013.

[21] S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *Proc. of WWW'09*, pages 381–390, 2009.

[22] M. Halbwachs. *On collective memory*. The University of Chicago Press, Chicago, 1950 (Translation).

[23] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. *Advances in Large Margin Classifiers*, pages 115–132, 2000.

[24] W. Hirst, E. Phelps, R. Buckner, A. Budson, A. Cuc, J. Gabrieli, M. Johnson, C. Lustig, K. Lyle, M. Mather, R. Meksin, K. Mitchell, K. N. Ochsner, D. Schacter, J. Simons, and C. Valdya. Long-term memory for the terrorist attack of september 11: Flashbulb memories, event memories, and the factors that influence their retention. *Journal of Experimental Psychology: General*, 138(2):161–76, 2009.

[25] D. Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, 2011.

[26] B. Keegan, D. Gergle, and N. Contractor. Hot off the wiki - structures and dynamics of Wikipedia's coverage of breaking news events. *American Behavioral Scientist*, 57(5):595–622, 2013.

[27] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in twitter. In *Proceedings of WWW '12*, 2012.

[28] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee. TwiNER: named entity recognition in targeted twitter stream. In *SIGIR*, 2012.

[29] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[30] D. Milne and I. H. Witten. An open-source toolkit for mining wikipedia. *Artificial Intelligence*, 2012.

[31] S. Mitra, M. Winslett, and W. W. Hsu. Query-based partitioning of documents and indexes for information lifecycle management. SIGMOD '08, pages 623–636.

[32] K. D. Naini and I. S. Altingovde. Exploiting result diversification methods for feature selection in learning to rank. In *ECIR*, 2014. to appear.

[33] K. D. Naini, R. Kawase, N. Kanhabua, and C. Niederee. Identifying high-impact features for content retention in social web applications. In *WWW (Companion Volume)*, 2014. to appear.

[34] T. N. Nguyen and N. Kanhabua. Leveraging dynamic query subtopics for time-aware search result diversification. In *Proceedings of ECIR '14*, 2014.

[35] M. Osborne, S. Petrovic, R. McCreadie, C. Macdonald, and I. Ounis. Bieber no more: First story detection using Twitter and Wikipedia. In *SIGIR TAIA Workshop*, 2012.

[36] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia, 1998.

[37] M.-H. Peetz and M. de Rijke. Cognitive temporal document priors. In C. Eickhoff and A. P. de Vries, editors, *DIR*, volume 986 of *CEUR Workshop Proceedings*, pages 36–37. CEUR-WS.org, 2013.

[38] C. Pentzold. Fixing the floating gap: The online encyclopaedia wikipedia as a global memory place. *Memory Studies*, 2(2):255–272, 2009.

[39] D. Rafiei, K. Bharat, and A. Shukla. Diversifying web search results. In *Proceedings of WWW '10*, 2010.

[40] D. Rafiei, K. Bharat, and A. Shukla. Diversifying web search results. In *Proc. of WWW'10*, pages 781–790, 2010.

[41] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of WWW '10*, 2010.

[42] R. L. T. Santos, P. Castells, I. S. Altingovde, and F. Can. Diversity and novelty in information retrieval. In *Proc. of SIGIR'13*, page 1130, 2013.

[43] C. A. Soules and G. R. Ganger. Connections: using context to enhance file search. In *ACM SIGOPS Operating Systems Review*, volume 39, pages 119–132. ACM, 2005.

[44] F. M. Suchanek. *Automated Construction and Growth of a Large Ontology*. Doctoral dissertation, Universität des Saarlandes, 2009.

[45] O. Tsur and A. Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *WSDM*, 2012.

[46] M. R. Vieira, H. L. Razente, M. C. N. Barioni, M. Hadjieleftheriou, D. Srivastava, C. T. Jr., and V. J. Tsotras. On query result diversification. In *Proc. of ICDE'11*, pages 1163–1174, 2011.

[47] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *Proc. of SIGIR*, pages 115–122, 2009.

[48] T. Yasseri, S. Anselm, M. Graham, and K. Janos. The most controversial topics in wikipedia: A multilingual and geographical analysis. In P. Fichman and N. Hara, editors, *Global Wikipedia: International and cross-cultural issues in online collaboration*. Scarecrow Press, 2014 (to appear).

[49] Z. Ma et. al. On predicting the popularity of newly emerging hashtags in Twitter. *JASIST*, 2013.