EEXCESS

Enhancing Europe's eXchange in Cultural Educational and Scientific reSources

Deliverable D5.1

Usage Pattern and Context Detection Specification and Analysis

Identifier:	EEXCESS-D5.1-Usage-Pattern-and-Context-Detection-Specification-and-Analysis- final.pdf
Deliverable number:	D5.1
Author(s) and company:	Christin Seifert (Uni Passau), Jörg Schlötterer (Uni Passau), Aenne Löhden (ZBW), Christina Niklaus (Uni Passau), Gerhard Doppler (BITM), Kim Plassmeier (ZBW), Henning Manske (ZBW)
Internal reviewers:	Know-Center
Work package / task:	WP5, Task 5.1, 5.2 and 5.3
Document status:	Final
Confidentiality:	Public
Version	2013-12-14

History

Version	Date	Reason of change
1	2013-11-29	First Draft Created
2	2013-12-06	Integrated contributions from ZBW
3	2013-12-10	Integrated contributions and review comments from WP6
4	2013-12-18	Integrated internal review comments
5	2013-12-19	Final Version

Impressum

Full project title: Enhancing Europe's eXchange in Cultural Educational and Scientific reSources
Grant Agreement No: 600601
Workpackage Leader: Christin Seifert, Uni Passau
Project Co-ordinator: Silvia Russegger, JR-DIG
Scientific Project Leader: Michael Granitzer, Uni Passau

Acknowledgement: The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 600601.

Contents

1 Executive Summary

This document summarizes the work in Work Package 5 – User and Usage Mining. The first goal of this work package is to provide all information from users to ensure good personalized recommendation results. The second goal is to collect usage data from EEXCESS resources in order to enhance recommendation results by improving resource description and incorporating statistical knowledge.

This document provides details on

- The EEXCESS User Profile Definition (also termed user model in other research fields)
- A first analysis on user mining algorithms and possible results and an estimate of the learning complexity.
- A first analysis on usage mining focusing on online social media channels.
- Privacy considerations with respect to the user profile and data collection.

The outcomes of the first analysis are the following:

EEXCESS User Profile and Mining The user profile contains long-term and short-term profile information capturing (among others) user interests, basic demographic data and relation to resources. With user context we mean non-aggregated information about the user or her surroundings, e.g. the location of the user and the user' focus within a web-page. We provide an estimate of the learning complexity of the different parts of the user profile as well as means to integrate short- and long-term information. Based on the complexity estimates we will research topic of interest and task detection methods next.

Test Data Acquisition To train and evaluate the machine learning models we identified the needs to collect test data. We developed a detailed test data acquistion procedure and respective user interfaces. The test data collection with multiple users is currently in progress.

Usage Mining Based on the general architecture, and the availability of information of different levels of the architecture the usage mining components will be distributed. For instance, parts of the usage mining can only be done with high-quality on the level of the privacy-proxy. We will first research mining of usage information from online social network, more specifically Twitter.

Privacy The users interests are modeled hierarchically. With respect to privacy considerations for users interests we designed an experiment to find out on which level of the hierarchy users are hidden within a group with their interest profile (k-anonymity).

2 Introduction

2.1 Purpose of this Document

The deliverable specifies the details on extracted user profiles, patterns to be detected and provides a first analysis of achievable accuracies/efficiencies.

2.2 Scope of this Document

This document covers the conceptual ideas of user and usage mining, related work and prototypical implementations. Covered by deliverable D1.1 [D11, 2013] and not part of this deliverable (D5.1) are the following aspects:

- User and usage mining components architecture
- Integration of components in the EEXCESSS architecture
- Mining components' technological state-of-the art
- The query flow inside the EEXCESS system (how a query in the back-end partner systems is generated from a user's information need, the user profile and the context)

Covered by deliverable D6.1 [D61, 2013] and not part of this deliverable (D5.1) are the following aspects:

- Privacy consideration and possible attack variants
- Argumentation for design choices on how the user and usage mining components interact with the other components

2.3 Status of this Document

The status of the document is draft with Christin Seifert being the responsible person. Some comments (colored bubbles) are for internal communication only.

2.4 Related Documents

Before reading this document it is recommended to be familiar with the following documents:

D1.1 First Conceptual Architecture and Requirements Definition [D11, 2013]
In detail the following sections are of interest for the reader:

- The architecture overview in Section 3.1., specifically the architecture of the client component in Section 3.2.1
- The querying workflow in Section 3.3.
- Technological state-of-the-art of WP 5 in Section 3.4.5
- Specific requirements derived for user and usage mining components in Section 4.4.

D4.1 Integration and Enrichment Specification Analysis [D41, 2013]

- Approaches to apply usage mining results for social enrichment are covered in Section 5.4

D6.1 Policy Model for Privacy Preservation and Feasibility Report [D61, 2013]
In detail the following sections are of interest for the reader:

- Current architectural choices from the privacy perspective in Section “EEXCESS context for privacy”
- Trustworthyness of the user and usage mining component for the first prototype in Section “Current results of WP6 - Impacts of different trust scenarios”

3 User and Usage Mining

3.1 A General Model for Users and Resources

In a very simplified manner, in this work package we want to know everything about users (user mining) and everything about resources (usage mining) and how users and resources are connected. This can be modeled as a graph, having two different types of nodes and relations between the nodes. One node type corresponds to users and the other corresponds to resources. Furthermore, resources and users reside in different channels. Such channels are the original repositories of resources (the data providers web sites), social networks or other other information hubs. The channels therefore act as mediator between users and resources and form a third type of nodes in the graph.

Users and resources are connected by semantic relations. Such relations are for instance: users like, share, view, annotate, bookmark a resource (and correspondingly a resource has been liked, shared, viewed, annotate or bookmarked by a user). This graph is not tripartite graph, since there can exist relations between different users and between different resources. Relations between users are for instance connections in a social network. Relations between resources might be that two resources belong to the same collection, the same topic, or more generally are semantically related. A relation between a resource and a user can only exist if there is a relation between a resource and a channel and a user and the same channel. This means, the users has to have access to a channel the resource has been published in [1]. Figure 1 depicts this model for users, channels and resources.

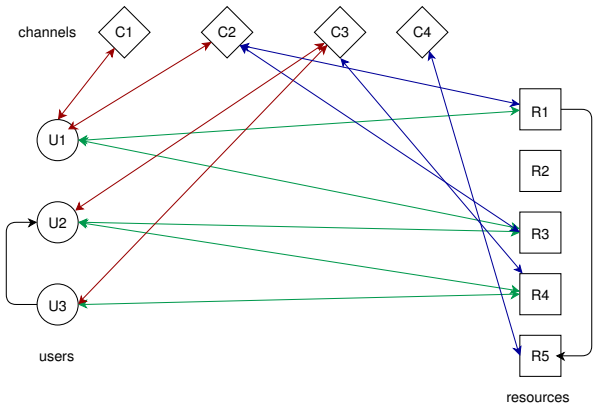

Figure 1: A user and usage mining model. The graph consists of three types of nodes, users, resources and channels. Connections between users and resources can only exist if users and resources reside in the same channel.

Basically, the graph can be viewed from two different perspectives: from the perspective of a user or from the perspective of a resource, both are detailed in the next two sections.

3.2 User-centric view

In the **user-centric view**, all information for one user are collected, i.e., all relations from the user to related resources and other users. Additionally other information may be added, such as preferences, interests, demographic information, physical context. This information can then be used to generate high-quality recommendations tailored to the users need. To mine the user-centric data we aim for a mining component installed in the browser. Thus, the mining component can access all user-centric data – as long as the user's privacy settings allow it.

[1] To establish the connection between users and resources from long-tail domains is exactly the goal of the EEXCESS project

3.3 Resource-centric view

In the **resource-centric view** all information about a resource are collected, i.e., which users are in which ways related to this resources, in which channels has this resource been distributed, how often has it been rated and how. This information can then be integrated to the resource provider's repository in order to enhance the quality of the recommendation service. To mine the usage data we aim for different stand-alone mining components for each channel.

3.4 Mining components

We use different components for user and usage mining for the following reasons:

- The browser-plugin will not be installed by all potential users of the resources, thus we can not collect all data that is required via the plugin.
- Different channels have different access policies and APIs, and new potential channels may emerge over time.
- The final EEXCESS privacy-policy may not allow to collect all necessary information on the client side and transfer it to the server.
- Pure client-side mining would not give the full picture of resource usage, as partner servers are also accessed directly or via different channels than the EEXCESS services.
- Pure partner server-side mining would not give the full picture, as we can not infer (i) which resources were accessed through the EEXCESS services, and (ii) privacy-settings may hinder to relate resources to users or user groups.

As a single mining component, whether on server or client side will not give us an exhaustive view on the user-resource relations, we aim for mining components on different levels of the architecture (cf. deliverable [D11, 2013]., section 3.2.6). Table 1 provides an overview of which information are collected by which component.

CUR Client-side user and usage mining (Client-side User and Resource mining (CUR)): Detailed user information and usage statistic of resources is collected at the client,

PPM Privacy-proxy usage mining (Privacy-Procy mining (PPM)): Collection of aggregated statistics over all users, e.g. search queries, delivered resources. Those statistics are only available in the privay-proxy, because i) the client holds only the profile for a single user and ii) the recommender might not know the real user queries due to query modification and also does not know which resources were finally delivered to the user when the privacy proxy applies filtering.

ESR EEXCESS server-side usage mining (EEXCESS Server-side Resource mining (ESR)): Social media channels, in particular Twitter, is monitored also for usage of a subset of resources.

PSR Usage mining on partner servers (Partner Server-side Resource mining (PSR))

Table 1: Overview of user and resource information collected by components.

information	examples	comp.
Usage statistics of resources accessed through EEXCESS client services[1]	user has rated a resource, user has clicked on a resource, user has added the resource to personal bookmarks, time user has spent on a specific resource	CUR
General user interactions with EEXCESS client services for single user[1]	topics of websites the user has visited, search queries and topics of the search queries the user has issued, focus area of the current website	CUR
General hardware and context information of EEXCESS client installation[1]	browser profile, geo-location	CUR
Personal (demographic) information about users[1]	manually input through the client: age, gender, interests, nationality	CUR
Aggregated user and usage statistics.[1]	keyword statistics, recommended items statistics, user groups	PPM
Resource usage in external (social media) channels[4]	how often has a resource been shared, comments on a resource	ESR
Partner-server log file information[2,3]	time-stamp of accessed resources, optionally a valid IP address, search queries (possibly anonymized)	PSR

[1] may not be allowed to leave the client due to the privacy policy
[2] no information whether users are EEXCESS users
[3] no information about whether the search query has been perturbed or is original
[4] only aggregated information over users available, since unique users can not necessarily identified in different channels

3.5 Summary

EEXCESS's first approach to personalized recommendations is query modification based on the user profile. Detected user interests and tasks will be used to reformulate the query accordingly. In the next section (section 4) we describe the user-centric view starting with the definition of a user profile, describing our first approach to automatically detecting the user profile and a first prototype. Section 4 covers the results of tasks 5.1 and 5.2. of the DoW. Section 5 focuses on integration aspects with respect to privacy considerations and on the anomyization of the user profile covering results from task 5.2. The resource-centric view is described in section 6 focusing on analyzing online social networks, in particular Twitter. This section describes the outcome of task 5.3.

4 User Profile and Context Learning

This section reviews work that has been done within WP5, tasks 5.1 and task 5.2. Work that is very specific with respect to the privacy aspects is reported in section 5.

First we define the EEXCESS user profile, consisting of long-term information, short-term information and contextual information in section 4.1. In section 4.2 we review approaches of learning the user profile. Also part of this section is the aquisition of test data necessary for supervised learning approaches. The interplay between privacy-aspects, the recommender and the user profile is described in section 5.1. The first Web-based user mining prototype is outlined in section 4.4, a conceptual idea for a prototype on mobile devices is described in section 4.5. Section 4.6 then summarizes the chapter and outlines related research questions and the approaches for project year 1.

4.1 EEXCESS User Profile

A user profile is a machine-processable representation of a user model for the purpose of user identification and personalization [Carberry et al., 2013]. With a user profile, each user is given an unique identity within the computing system and can be identified by the profile information. The main purpose of user profile is to make systems adaptable to personal users' needs by "Saying the 'right" thing at the 'right" time in the 'right' way."[Fischer, 2001], i.e., to personalize the way the user interacts with the computer.

Within the EEXCESS project, this personalization refers to recommendations and visualizations. Each user should get his or her personalized list of recommendations, and the visualizations should be adapted to the users needs. Within the project we have further the requirement of preserving the privacy of the user. This means, that no information that may reveal the user of information about the user may be transferred across services.[2] For the purpose of this paper we assume that the anonymization process is a black-box operating on a user profile before it is transferred. Figure 2 shows an overview of that process, where the anonymization process is called "privacy-preserving proxy". What we describe in the following is what is called "explicit user profile" in the figure.

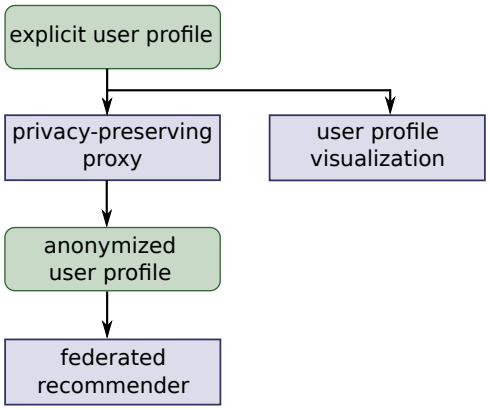

Figure 2: Explicit and anonymized user profile. Explicit user profile is described in this paper. Only the anonymized user profile is transferred to other services. Visualization is performed on the client-side on the explicit user profile.

[2] Depending on the outcome of WP 6, users might be inclined to transfer a subset of profile information non-anonymously. But until we know for sure we go with the stricter assumption.

4.1.1 Profile Information vs. Context

Depending on the research field implicit or explicit information about users is also termed "context" (see for instance [Dourish, 2004]). Baziere and Brézillon state (after their analysis of 150 different notions of context), that "a definition of context depends on the field of knowledge that it belongs to"[Bazire and Brézillon, 2005] We use the term *user profile* for all information about a user, and differentiate between *long-term* and *short-term profile information*. We refer to the term *context* as additional information, which is relevant to the situation at hand. In our notion, context has a representational role that can be captured at the current state and does not refer to general background, including experience, knowledge and beliefs. As just mentioned, the information in the user profile can be distinguished into *long-term* and *short-term* information. For instance, while the year of birth of a user remains constant over time, the users interests may shift depending on the occupation (starting a new job) or even the task at hand [Li et al., 2007, Bennett et al., 2012]. We reflect this by the use of a *long-term* and a *short-term* profile. The former contains characteristics of the user, which remain constant over large intervals, adapt only slowly over time or do not change at all. The latter is a representation of the user's characteristics and actions within a recent time span. Figure 3 illustrates how the user model is composed by the *long-term* profile, *short-term* profile and *context*. As some

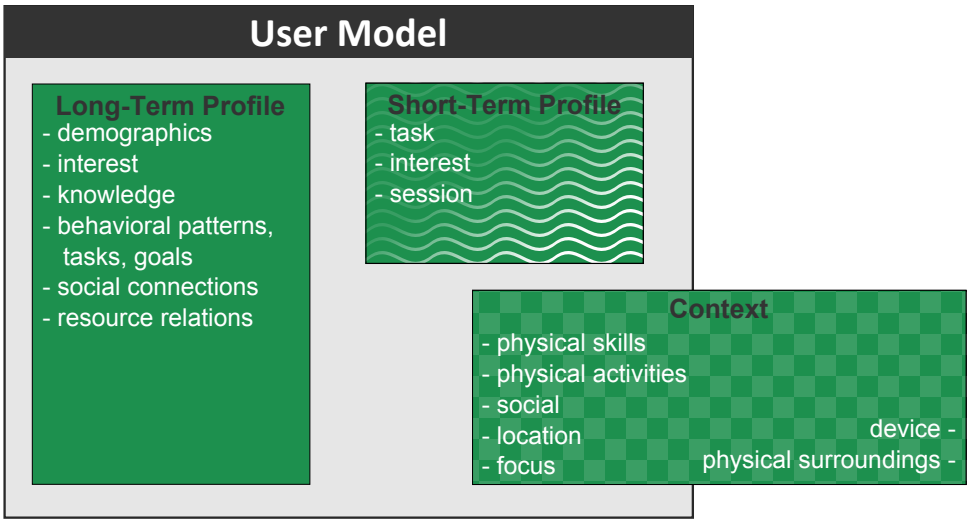

Figure 3: EEXCESS user model - composition of long-term profile, short-term profile and context

contextual features, such as device capabilities are not part of the user model itself, but nevertheless an important source of information about the user's surroundings, they are not contained within the user model, but coupled to it. A brief outline of long- and short-term profile information and context dimensions contained in our user model and shown in the figure is provided subsequently.

The EEXCESS user model consists of the following *long-term information*, which are explained in detail in section 4.1.2:

- **Demographics:** Demographical information (e.g., name, birthday, education level).
- **Interest:** General topics of interest for a user and a weight for each topic .
- **Knowledge:** Topics for which the user is knowledgeable about, a weight models the user's degree of knowledge about a particular topic.
- **Behavioural Patterns, Tasks and Goals:** High-level behavioral patterns and typical high-level tasks of the user, e.g. "shopping online for something", where it does not matter what the specific "something" is.

- **Social Connections:** Relations to groups or others in social networks.
- **Resource Relations:** Relations to resources of interest, e.g. resources from the recommender data base.

The EEXCESS user model consists of the following *short-term information*, which are explained in detail in section 4.1.3:

- **Task:** Current user task, which is one of the typical task from the long-term profile or "undefined".
- **Interest:** Topics of interest for the current task, a subset of the topics from the long-term profile or "undefined".
- **Session:** A session is a sequence of topics and tasks, i.e. "searching the web for literature on Recommender Systems" could be a session having the tasks "Find conferences", "Search ACM digital library catalog".

The EEXCESS user model consists of the following *context information*, which are explained in detail in section 4.1.4 (as context depends on its belonging field, this list can never be comprehensive and context in a particular application area tends to be a subset of it though):

- **Physical...**
  - **Surroundings:** Environmental conditions, e.g. weather
  - **Skills:** The user's physical skills
  - **Activity:** The activity currently performed by the user
- **Social Surroundings:** Social setting, e.g. whether the user is with family or colleagues
- **Device Capabilities:** Capabilities of the interacting device, e.g. screen resolution, available plugins
- **Location:** Geographical information, such as latitude & longitude coordinates, elevation, proximity
- **User Focus:** The entity in the user's most narrow focus

4.1.2 Long-Term Profile Information

The long term profile contains rather static user attributes, i.e. properties, which remain (fairly) constant over large intervals. For example, a user's birthday or birthplace do not change at all, while his name may change, but only at rare occasions, such as getting married. Thus, the long-term profile information can be seen as a description of the user in general, aggregating (in an ideal view) the user's whole life, up to the current point of time. The following dimensions are model in our definition of the user profile:

Demographics This dimension contains demographic attributes, such as name, birthday, profession and education level (e.g, A-Level). According to Brusilovsky et al., attributes like profession are "nearly impossible to deduce by simply watching the user work" [Brusilovsky and Millán, 2007] and thus typically provided explicitly. Nevertheless, some of the demographic features may be mined implicitly, for example the user's address (given the assumption of always being provided with the user's current location and equating the user's address with the location she stays most of the time). Beyond that, Kosinski et al. were able to predict personal attributes like age or gender on the basis of facebook likes [Kosinski et al., 2013]

For the attributes *first name* and *last name* from the demographical features, the *Friend of a Friend* (foaf[3]) vocabulary is used and for the other demographical information we use the *General User Model Ontology* (*gumo* [Heckmann et al., 2005]). For more information on the latter see the appendix.

Interest Interests in the long-term profile information are not totally constant, but adapt only slightly over time. They are represented by the topic of interest and an associated weight, reflecting the level of interest. This weight decreases over time, if no evidence is seen for a particular topic and gets increased on the opposite. The topics are represented by dbpedia categories and resources. Regarding categories, this representation reflects a kind of hierarchy via the *Simple Knowledge Organization System* (*skos*[4]) relation *broader (of)*, which is not strict, but according to the Wikipedia categorization[5] more like a directed acyclic graph (although cycles are undesirable, they occur in the Wikipedia hierarchy). Resources constitute the leaves in this hierarchy and are mapped to the corresponding concepts via *dcterms:subject*. Evidence seen for a topic on a fine grained level will bubble up and increase the weight at its parent topics accordingly. The hierarchical approach enables to deal with situations, requiring information on a very fine grained level, as well as with situations, that demand only broad information. For example, if very specific resources are available at a data provider for a certain field, fine grained information will help to reveal the desired results, while at fields with less specific resources a search needs to be based on a more coarse level to provide results at all. Moreover, this approach enables the adjustment of individual levels of privacy by restricting the access to the user's interest to a certain layer in the hierarchy (cf. Section 5.2).

To model the interests, we use the *Weighted Interest Vocabulary* (*wi*[6]) and *Weighting Ontology* (*wo*[7]). They provide the ability to assign weights of interest to particular topics, along with dynamics, such as temporal ones (e.g. a topic may be relevant only at particular times).

Knowledge For the knowledge dimension, we use the same concept as described for the user's interests in the previous section 4.1.2. Languages, which are not the user's mother tongue are also contained in the knowledge dimension, modeled as topics with a certain knowledge level, as well as locations, the user has visited or lived in. This choice was made because having lived in a city implies a certain degree of knowledge about it, whereupon this degree of knowledge is the valuable information, rather than the fact that the user has lived in this city. Hence, this information is transformed from the location domain to the knowledge domain instead of introducing a separate dimension to the user profile. This applies only for the long-term profile - the user's current location is included in the context, as it is a valuable source of information in some applications.

Since the features, required to describe the user's knowledge about a particular topic are the same as for describing a user's interest, we apply the same vocabularies to the knowledge domain as to the interest domain: we use an own subclass of the *WeightedInterest*-class from *wo*, which we call *WeightedKnowledge*, with its name being the only difference at the current point of time. Topics are again represented by Dbpedia categories and resources.

Behavioral Patterns / Tasks / Goals High-level patterns exhibited by the user over a longer time period are modeled in this dimension. The goals correspond to specific tasks in this dimension and do not reflect general ones like "world peace" for example. In the first approach, we will solely focus on tasks. We refer to tasks as templates of small work units, whose actual occurrences are a combination of the template with related topics. An example of such a task is "searching the

[3] http://xmlns.com/foaf/spec/
[4] http://www.w3.org/TR/skos-reference/
[5] http://en.wikipedia.org/wiki/Wikipedia:Category
[6] http://purl.org/ontology/wi/core
[7] http://purl.org/ontology/wo/core

ACM-Portal[8] for scientific literature [task template] on probabilistic models [related topic]" with the goal of finding relevant articles. These tasks can be further decomposed into almost arbitrarily many sub-tasks, depending on the level of granularity chosen (e.g. entering a query term into the search box -> entering a single character -> hitting the particular key on your keyboard and so forth). The tasks represented in the user profile range around the abstraction level given by the example though.

The representation of tasks is based on the *Task Model Ontology* (*tmo*[9]) and extended by the frequency, the user executes a particular task, in order to distinguish commonly performed tasks from those, which occur only rarely.

Social Connections The user's connections in social networks are modeled within this part of the long term profile. For each connection, attributes, describing the ties present and the type of the connection are provided. We use a mix of the *Semantically-Interlinked Online Communities* (*sioc*[10]) ontology and the *foaf* vocabulary to represent the user's social network.

Resource Relations This dimension contains resources, which have been recommended to the user as well as resources, the user has viewed without an explicit recommendation or interacted with in some other kind of way. The resources are attributed with a timestamp of when the action occurred, an annotation about the resource (for a recommendation, this may be a rating for example) and a flag, indicating whether the resource has been recommended or not. For the representation in the user profile, we chose the *Open Annotation* specification (*oa*[11]), which considers an annotation to be a set of connected resources, typically a body and a target where the body is somehow about the target. Thus, an annotation can be used to provide a rating for a resource, which has been recommended to the user, or can reflect a user's assignment of a label to a bookmarked webpage as well. In addition, the resource relations contain the user's publications (if any), modeled with the *foaf* vocabulary, since those are of outstanding interest and may serve to deduce interests.

Table 2 provides an overview of the long-term user profile's dimensions and attributes, along with their expected impact on recommendations, expected complexity on mining and the vocabulary used. An exemplary long-term profile for the fictive user "Horst B." can be found in the appendix A.1, along with its visualization.

At mining the long-term profile, we start at features, which we expect to have a high impact on recommendation quality (and on the question, of when recommendations should be provided at all) and low learning complexity (those would be marked green/green in table 2) and iteratively move on to the features from which we expect a lower impact, tackling attributes with low impact and high learning complexity at last. Thus, we focus on the user's interests at first to deduce her preferences for recommendations and on the user's tasks, to determine at which points in time providing recommendations is beneficial and at which points in time it would not help or even annoy the user and should therefore be avoided.

4.1.3 Short-Term Profile Information

The short-term profile represents information about the user that is restricted to a certain time span, reflecting the user in the most recent past. Hence, the short-term profile is kind of a sliding window and its contents change frequently with the progress of time. Within a web context, parts of the short time profile may be derived from the browsing history, utilizing for example the ten websites visited at last.

[8] http://dl.acm.org/
[9] http://www.semanticdesktop.org/ontologies/2008/05/20/tmo
[10] http://rdfs.org/sioc/spec/
[11] http://www.w3.org/ns/oa

Table 2: Dimensions and attributes of the user profile

attribute	im-/explicit	expected impact	learning complexity	vocabulary
Interest				
topic	I	H	M	**skos**
weight	I	H	M	**wi,wo**
Knowledge				
topic	I	M	H	**skos**
weight	I	M	H	**wi,wo**
Demographics				
profession	E	M	H	**-**
education level	E	M	H	**-**
institution	E	M	H	**-**
first name	E	L	H	**foaf**
last name	E	L	H	**foaf**
birthday	E	L	H	**gumo**
birthplace	E	L	H	**gumo**
address	E	L	L	**gumo**
- city				
- country				
- house-nr				
- postal code				
- state				
- street				
Social Connections				
connections in social networks	I	M	M	**sioc/foaf**
- strong/weak ties				
- type of connection (groups)				
Resource Relations				
resource	I	M	M	**oa/foaf**
- timestamp				
- annotation				
- been recommended				
Behavioural Patterns / Tasks / Goals				
tasks	I	H	H	**tmo**

expected impact - Low (L), Medium (M), High (H)
the expected influence of a particular attribute on recommendation quality

learning complexity - Low (L), Medium (M), High (H)
the expected complexity to learn a particular feature of the user profile automatically

im-/explicit - Implicit (I), Explicit (E)
whether the feature has to be given by the user explicitly or can be mined implicitly (the user may also change implicitly mined features manually)

Interest The interests in the short-term profile are a subset of the long-term profile's interests. Thus, they are represented with the same vocabulary as in the long-term profile. The weight of an interest in the short-term profile may be totally different from the weight of the corresponding topic in the long-term profile: the user may have high interest in a particular topic for a limited time period, which is absolutely not relevant to her in general, for example when she is shopping for a present for her daughter. Another example for highly weighted interests in the short-term profile, whose weights are not reflected in the long-term profile, is the occurrence of current events, drawing attention on certain topics only for a limited amount of time. But in general, the interest's weights of the short-term profile contribute to the weights in the long-term profile: a topic, frequently occurring with high weight in the short term profile, will produce a high weight value in the long-term profile as well.

Tasks Analogous to interests, tasks in the short-term profile constitute a subset of the long term profile's tasks. Therefore, they are modeled with the same vocabulary. In contrast to the long-term profile, a task in the short-term profile may additionally be "undefined", in the case of a previously unseen or not identifiable task. The combination of topics (which are specific) and tasks (which are templates) then define a specific user task. An exemplary task for the goal of finding literature about privacy in recommender systems is "literature search" with the relevant topics "recommender system" (`http://dbpedia.org/page/Recommender_system`) and "privacy" (`http://dbpedia.org/page/Privacy`). The weights of the two topics depend on the user's previous actions, but in this case, an equal distribution would be reasonable. ı+++tem[Session] A Session consists of a sequence of tasks and topics. While the tasks, as represented in the user profile, can be decomposed into smaller sub-tasks, the opposite holds as well. Therefore, a session is identical to an aggregation of tasks to a single super-task and hence is a task again. Moreover, a session may also consist of a single task only and in this case, session and task are absolutely equal in terms of representation, but used with different semantics. To emphasize the distinction between these two, we use the term session for the aggregated type. The need for recommendations by EEXCESS is derived on the basis of sessions. Hence, session-breaks are of high interest, since they indicate the points in time, when exactly this need changes.

4.1.4 Context

When talking about context, most people share a tacit intuition of its meaning that is hardly tangibly, when trying to express it explicitly. As already mentioned, the definition of context depends on its particular field and coherently, a whole bunch of definitions exists for context. A quite general definition of context that most closely fits our needs is the one given by [Dey et al., 2001]. They define context as: "any information that can be used to characterize the situation of entities (i.e., whether a person, place, or object) that are considered relevant to the interaction between a user and an application, including the user and the application themselves. Context is typically the location, identity, and state of people, groups, and computational and physical objects."

To distinguish precisely between profile and context, the profile in our notion is aggregated information, whereas the context is raw (sensor) data, or not observable at all (e.g. social connections or setting - whether the user is with her family or colleagues). An example for this distinction is the user's geolocation vs. her (implicitly deduced) address. Both are represented by geo-coordinates, but the first represents the user's current position (and thus is raw data), whereas the latter is aggregated by a set of geolocations (e.g. mined through movement patterns). By omitting imperceptible contextual features (since they can not be observed, we cannot make use of them), the typical representatives of context dimensions still stay the same as provided with the definition of [Dey et al., 2001]. Nevertheless, context is not limited to the given examples and, depending on the situation, may exhibit totally different features. One feature, we account to be part of the context is the user's most narrow focus.

In a web-setting, this is the paragraph of a web page, the user is currently looking at. An exemplary scenario for this is a user searching for literature on machine learning. Thus, the task is "searching for literature", and the interests are given by machine learning topics, deduced from the previously

visited pages. The short-term profile contains exactly those interests and they are also present in the long-term profile, whereas the latter typically holds information about other interests (and they interest's weights in the long-term profile may be smaller). An irrelevant feature in this situation are physical surroundings, such as weather and can therefore be ignored. Using the focus as a contextual feature may significantly influence the query terms: If the center of attraction (namely the viewed paragraph) would not be modeled explicitly, the query would be based solely on the long- and short-term profile, which indirectly incorporates the paragraph as well, but does not account for its special meaning. Thus, the query would be a more general one.

An alternative approach would have been to model the focus as a dimension of its own in the user model, since it might be considered as a central feature of the situation itself and not as additional information to it. But reconsidering the given example, we regard the task "searching for literature" as the central point of the situation to which the page focus only provides an additional contribution (even though it is a large one). Thus modeling the focus as a contextual feature seems more appropriate.

4.2 Learning the User Profile

In the following sections, we outline related approaches to deduce our user model implicitly from watching a user work and specify our plan for test data acquisition. A test data set is necessary to obtain labeled features for supervised learning and to have a baseline, against which we can compare implicitly mined models. To the best of our knowledge, no test data set, fitting our particular domain of interest is available so far. Thus, we aim to create a gold standard for the cultural heritage domain in addition.

4.2.1 Learning Approaches

This section presents already existing learning approaches for the particular dimensions contained in our user model and their interplay. Where available, accuracies and improvements over existing baselines are presented, giving a first intuition on how well the system could perform.

Long-Term Profile A long-term profile of term frequencies, constructed from the browsing history is used by [Matsuo, 2003] to determine relevant words in the current document. Their relevance is measured by the so called "Interest Relevance Measure" (*IRM*), represented by the biases of co-occurrence between a word in the document and words in the long-term profile. This measurement is similar to *tfidf* in terms of assigning small weights to words, which occur rather rarely in a document and those, which occur frequently in the document, but also in the user's long-term profile. The main advantage of *IRM* is the assignment of small relevance to words, which occur frequently in a document, but are irrelevant to the user's interests (i.e. they do not co-occur with any familiar words), whereas *tfidf* would assign higher relevance. Also based on the browsing history and the contents of visited pages for personalization, [Matthijs and Radlinski, 2011] were able to outperform the result ranking, provided by Google by around 20%. They point out that "the key to using web pages to model users is to not treat them as flat documents, rather as structured documents from which several types of data can be extracted". With the use of probabilistic models, [Sontag et al., 2012] gain an overall improvement at best of around 0.02 in mean reciprocal rank, compared to the baseline ranking provided by Microsoft Bing search. In the work of [Calegari and Pasi, 2013], the representation of the long-term profile is based on the YAGO[12] ontology. Their profile building process starts with a set of interest terms as seeds, extracted from documents, representing a user's particular interest topic and weighted with a tf-idf scheme. Based on these interest terms and synonyms derived from WordNet[13] corresponding entities are identified in YAGO. In the second step, these entities are disambiguated, in order to remove noise. A set of handcrafted rules is applied to the remaining set of entities, to extract

[12] http://www.mpi-inf.mpg.de/yago-naga/yago/
[13] http://wordnet.princeton.edu/

the relevant knowledge related to them from the ontology, which constitutes the long-term profile. The profile is evaluated with an adaption of the robustness index - (#relevant entities - #irrelevant entities) / #all entities, reporting values (were relevance is judged by the user), ranging from 0.6 to 1, for ten users.

Short-Term Profile [Turnina, 2013] proposes to model the short-term profile with a semantic network of concepts and terms, in order to be used for query expansion. The terms in the semantic network are given by search query terms and corresponding concepts are derived from an ontology. Terms are assigned decaying weights, corresponding to their occurrence frequency in queries and bubbling up to the corresponding concept. Weighted links between concepts are established, if terms of the different concepts co-occur in a query. These links can then be utilized to expand a query with terms from a related concept. Using a concept-based short-term profile as well, with session boundaries defined by the conceptual correlation between the profile and a current query, [Daoud et al., 2009] were able to increase average precision and recall over the baseline of top ten results by around 66% and 188% respectively.

Session & Task Using hierarchical agglomerative clustering, [Murray et al., 2007] claim to achieve 98% precision in identifying session breaks, compared to human judgments on an evaluation set, extracted from AOL search logs. Their analysis is based on user activity, not specifying a static threshold for timeout, but instead utilizing a user specific clustering criterion. They define this criterion as the maximum ratio between the distance of the length of a search interval to the mean of interval lengths and the standard deviation. The length of the search interval, for which this ratio is maximal, is the used as timeout threshold.

The feasibility of detecting user tasks was demonstrated by [Rath et al., 2008], [Granitzer et al., 2009]. The authors were able to detect the execution of predefined tasks with an accuracy of around 80% and user-labeled task with an accuracy of around 70%.

Analyzing multitasking in the web, [Buzikashvili, 2006] has discovered that users perform multitasking in less than one percent of their search session and that even in multitasking sessions, they perform two tasks at most. Moreover, the two tasks are usually executed in an "enveloped" manner: one task is interrupted, the other task is executed and after its execution, the interrupted task is continued. We reflect this behavior in our test data acquisition described in section 4.2.2, by collecting single (consecutive) tasks.

Impact of Long- and Short-Term Profile The impact of long- and short-term profiles to search result relevance was studied by [Bennett et al., 2012]. They define *long-term* as all historic search interactions, up to the current session, *short-term* as the current session, *aggregated* as a combination of both and *union* as an aggregation of these three features in which the ranker learns how to combine them. Each of these features provides a gain in mean average precision. Moreover, their findings show, that historic data provides the highest benefit at the start of a new session, short-term information contributes more and more gains, as the session proceeds and that a combination of both outperforms using either alone (with the learned aggregation provided higher gains than a simple combination). Due to proprietary reasons, they report solely the gains in mean average precision over the baseline, provided by Microsoft Bing, not the absolute values. Advantages of combining short- and long-term profile are also pointed out by [Dou et al., 2007], while they highlight, that choosing the appropriate personalization strategy is a key factor, including the question, when results should be personalized and when they should not. The latter problem is addressed by [Teevan et al., 2008], who achieve a prediction accuracy of approximately 80%, when incorporating query history and result set features.

Dynamic Adaption of Long- and Short-Term Profiles [Li et al., 2007] present a scheme for the dynamic adaption of long- & short-tem profiles, used for result reranking. Their long-term profile reflects

user preferences as topics from the Google Directory Project[14] (GDP), associated with a weight, based on the number of clicks on pages from this topic. The short-term profile is a page history buffer of size 20. Adaption is performed by replacing the page with the smallest click count in the short-term profile with a previously unseen page and degrading the weight of its corresponding topic in the long-term profile. Search results are re-ranked half by their similarity to the long-term profile and half by PageRank. The similarity is measured as the smallest distance between the result's topic and a topic in the long-term profile, incorporating in addition the associated weight of the latter. The authors report an improvement in ranking quality of about 29% compared to the regular GDP search.

Contextual Features The learning possibilities of contextual features strongly depend on the indicators and sensor, exposed by the contextual features of a particular field of application. In the web setting on a desktop machine, we focus on geolocation and the user's most narrow focus (i.e. the currently viewed paragraph in a web page). For the geolocation context, we utilize the HTML5 Geolocation API[15], whose accuracy depends on the actual implementation and the sensors available. If a GPS-sensor is available, the location can be determined with very high precision. Otherwise, less accurate methods like WIFI, IP-based or cellular positioning have to be used. [Zandbergen, 2009] report an average median error of 8m for GPS positioning, 74m for WIFI positioning and 600m for cellular positioning. [Poese et al., 2011] showed, that GeoIP databases provide reliable information at country level at around 97%, but, compared with data from a large European ISP as ground truth, less than 20% deviated within a tens of km.

Regarding the currently viewed paragraph, the actual content has to be extracted from the currently visible contents in the viewport (i.e. for example removing navigational elements). Therefore, we developed a heuristic, which is based on the amount of white space characters between text passages, since navigational elements tend to be nested in a couple of DOM-nodes and thus expose several white space characters, when their textual content is retrieved. This heuristic performed very well in the cases seen so far, but requires an in-depth evaluation though.

For context detection in a mobile environment see section 4.5

Topic Detection Detecting the relevant topics of a document, also termed text categorization, since the document is assigned to one or more categories, can be achieved with machine learning techniques implicitly. In a comparison of experimental benchmarks on five versions of the Reuters collection, [Sebastiani, 2002] concludes boosting-based classifier committees, support vector machines, example-based methods and regression methods to perform best, with neural networks and and online linear classifiers performing only slightly worse. Rocchio and probabilistic naive bayes classifiers expose the worst performance in this comparison. The breakeven points of recall and precision vary between the version, but range from around 0.6 to around 0.9. However, the author warns to handle the results with caution, since they might be influenced by background conditions, such as preprocessing, indexing, dimensionality reduction, etc. In addition, other application contexts may exhibit different characteristics than the Reuters collection and thus leverage the performance of another learning approach, invalidating the presented ranking.

While the Reuters collection comprises approx. 100 categories, [Rubin et al., 2012] have shown, that generative models outperform discriminative methods in multi-label classification, as the amount of labels increases and label frequencies are skewed. The failure of hierarchical support vector machines on large datasets with many classes was shown by [Liu et al., 2005], reporting a macro F1 score of 0.12 on the 15th hierarchy level of the Yahoo! Directory[16] (their utilized subset comprised 132,199 categories and almost 800,000 documents). The models in [Rubin et al., 2012] are based on LDA [Blei et al., 2003], a model in which each document is represented by a mixture of set of underlying latent topics. They report to achieve an average precision of around 0.63 on a subset of the NYT

[14] The Google Directory Project provided a search with ranked results, based on the data of the Open Directory Project `http://www.dmoz.org/`, a human-edited directory of web pages. It was discontinued in 2011.

[15] `http://www.w3.org/TR/geolocation-API/`

[16] `http://dir.yahoo.com/`

corpus[17] (comprising 30,685 articles and over 4000 labels) with their approach, compared to around 0.49, achieved with tuned support vector machines. The possibility to combine LDA with concept hierarchies was also shown by [Chemudugunta et al., 2008] and applied to web page categorization. The dbpedia categories utilized in our user profile amount to almost 100,000 for the English Wikipedia, which contains around 4 million articles.

Reliability of Implicit Feedback Most of the aforementioned approaches rely on implicit feedback measures, such as clicks on recommended search results and the time spent viewing a particular item. It has been shown that display time is an indicator for the level of interest, when reading newspaper articles or browsing the web [Morita and Shinoda, 1994][Claypool et al., 2001], but reading time behavior is biased by the corresponding task and not suitable for relevance judgment [Kelly and Belkin, 2004]. While the latter findings are confirmed by [Kellar et al., 2004], their studies indicate, that usefulness of reading time increases, as the task becomes more complex (relevance judgment vs. simple question answering vs. complex question answering). This aligns with the results of [Fox et al., 2005], claiming reading time to be a helpful predictor of user satisfaction for search sessions. For clickthrough data, [Joachims et al., 2005] have discovered that interpreting them as absolute relevance measure is difficult, but reasonable accuracy (around 80%) can be achieved by deriving relative preferences, compared to the accuracy of agreement between explicit relevance judgments (around 86%).

Client- vs. Server-Side Retaining 97% of the performance of server-side profiles in an experimental evaluation, [Bilenko and Richardson, 2011] show that client-side profiles can compete with server-side profiles. In their work, they construct keyword user profiles for advertisement platforms, by adding keywords to the profile, which maximize its utility. However, an utility estimator still has to be computed server-side.

4.2.2 Test Data Acquisition

Ground truth data is required to enable the evaluation of implicitly mined features against their actual values. In our case, this is particularly relevant to determine the accuracy of extracted user profiles and context, as well as for the quality of personalized recommendation results. Hence, we need to gather test data, which provide real world examples of users' interactions and needs. The data we need to collect can be split into two dimensions basically: The behavior of a user and search quality. The latter describes not only the relevance of recommendations, but starts already with the query generation, based on the user's preferences. In the following, we describe the features of each dimension we identified to be required for the evaluation.

User Behavior First of all it is necessary to determine the points of time, when recommendations are desirable. Thus we need to identify the tasks at which a user has the need for additional information. Therefore, we need to know, which task a user is currently performing. This includes tasks, which are predominantly relevant to EEXCESS, for example "writing a blog" entry or "browsing for information" (since at those tasks recommendations are desirable), as well as tasks, which will not trigger EEXCESS-recommendations. Sampling both types of tasks provides the ability to distinguish between tasks, at which recommendations are desirable and at which they are not. Moreover it enables us to identify additional tasks, at which the users would benefit from recommendations by the EEXCESS framework.

Closely related to the collection of tasks, we need to identify session breaks. The session breaks constitute a change in the user's goals (e.g. switching from browsing the web for shopping a mobile phone to browsing the web for information about a particular historical artifact). Consequently these breaks delimit the spans of tasks at which recommendations should be provided or not.

[17] http://catalog.ldc.upenn.edu/LDC2008T19

For each task, we need to know the topics, that are related to this task. They form the basis for interests in the short-term profile and interests and knowledge in the long-term profile accordingly. With collecting these topics, we have a set of correct labels for a supervised learning approach, which will infer interests and knowledge based on the browsing history. Hereby, the topics need to be extracted from the content of the visited pages. Regarding interests, the duration time of a visit will influence the weight assigned to a topic, while for knowledge, this is a more complex task. One could argue, that the time spent with a certain subject correlates with the knowledge about it, but this does not capture previous knowledge about it. Even though a higher increase of knowledge about a certain topic is assumable, the longer the user engages with this topic, this is not necessarily true: a user may also read a text without understanding anything.

Nevertheless, the quality of recommendations also depends on the knowledge of a user about the topic of interest: An expert in a particular domain has the demand for very in-depth recommendations, while a beginner favors results on an introductory level. Thus, ratings for results may vary, depending on the background knowledge of the user about the particular topic. To reflect this in our test data set, we ask the users to give their level of expertise for each executed task. This level will also serve as label for a supervised learning approach, deducing the weight of knowledge.

In order to meet a user's privacy concerns, we need to identify, which personal information is considered highly sensitive, which is considered less sensitive and to which degree of sensitivity a user is willing to disclose information about herself. Knowledge about the information users are willing to disclose about themselves (demographic information), their location and their interests, provides the basis for privacy preservation mechanisms, hiding those parts of information, a user is not willing to provide publicly.

Summarizing the user behavior features of interest, we need to collect the following:

- the task a user is currently performing
- session breaks, at which a user is changing her goal
- topics related to the currently performed task
- level of expertise on the performed task
- disclosure level of personal information

Search Quality As mentioned in the previous paragraph, we need to know if recommendations are desirable at all. Thus, we need an indicator for each task that tells us whether we should recommend anything or not.

In order to be able to phrase queries that will fit the user's need most precisely, we need to obtain the search queries, a user would issue by herself and keep track of the query's context. We denote the query context as the user's most narrow focus related to the issued query. For instance, this may be a paragraph in a website. Given the knowledge of which queries were issued in which context, will enable us to deduce the query terms for a given context automatically.

The user's feedback is required to evaluate the quality of recommendations. This includes explicit feedback via a rating system, as well as implicit feedback. The most important fact in terms of the rating is whether a recommendation was helpful or not, thus we need to collect this information. For recommendations, the user has not rated explicitly, we might gather implicit feedback though. Viewing the resource of a recommendation constitutes a certain rate of interest at least and the duration of a view provides an informative basis of the resource's relevance.

Beyond this feedback on specific search queries and their results, we need the user's to asses if recommendations and the interface were helpful for the execution of the task at hand.

As the quality of recommendations will (amongst others) depend on the information available about a user, there is a tradeoff between providing high quality recommendations, based on a fully detailed

user profile and recommendations with lower quality, based on a subset of the full profile, due to privacy conservation mechanisms. We want to know how a user would deal with the impacts of privacy preservation. In particular, we are interested if a user is willing to disclose more information about herself, if she perceived a loss of quality after restricting the set of available information. If she is willing to do so, we want to know, which types of information she would disclose (highly sensitive vs. less sensitive).

The data to collect regarding search quality summarize as follows:

- indicator if recommendations are desirable
- issued search queries (and context)
- quality of recommendations
- assessment if recommendations & interface were helpful
- impacts of reduced quality on will to disclose personal information

Table 4 provides an overview of the features collected from the users along with the according methods of collection. The questionnaire with the respective questions on qualitative issues can be found in the appendix A.2.

Table 4: features collected with user tests along with respective methods

feature	**method**	im-/explicit
predefined task name	UI-control [select field]	E
custom task name	UI-control [input field]	E
task start-time	UI-control [button]	E
task end-time	UI-control [button]	E
level of expertise	UI-control [slider (range 0-10)]	E
topics relevant to task	UI-control [input field]	E
input language of topics	UI-control [select field]	E
indicator, if recommendations are desirable	UI-control [checkbox]	E
search queries	UI-control [input field]	E
rating of recommendations	UI-control [button (good/bad)]	E
assessment if recommendations & interface were helpful	question	E
assessment of sensitivity level of particular personal information	question	E
disclosure level of personal information (subject to the recommender's quality)	question	E
clicked recommendations	implicit	I
ignored recommendations	implicit	I
dwell time at recommendation preview	implicit	I
mouse clicks (+target)	implicit	I
textual input	implicit	I
browsing history	implicit	I
browser profile (plugins, ...)	implicit	I

Acquisition Procedure The acquisition of the test data set is aligned at the scenarios described in the deliverable D 1.1. We aim to address the two general tasks of content creation and content consumption. For both tasks, we use an adapted version of the browser extension prototype to collect the test data, together with a questionnaire to assess qualitative issues, such as interface concerns and the question, if recommendations were helpful. The questionnaire also serves to answer questions related to privacy, such as the amount of information a user is willing to disclose (see paragraph 4.2.2) and the amount of private information, a user is willing to trade for higher quality of recommendations (see paragraph 4.2.2).

The content consumption task is to annotate a given website with recommended resources. Hereby, the annotation is solely needed to collect the required data and thus the task is viewed as content consumption, since the task behind it is consuming a website. Annotating a website with recommended resources provides an assignment of resources to paragraphs or words and hence reveals, which resources are to be recommended as additional information, when a user reads through the website's contents. The users are given different websites to annotate with as much as relevant information as they see fit. As stated previously, the users have to phrase the queries for resources by themselves.

The content creation task is to write a blog entry about a given topic. The topics to write about are semi-defined: They comprise an important historical event, a cultural sight of the user's hometown (or another town of her choice) and a person, who played a significant role in history. The semi-defined topics provide the ability for the user to choose a topic, she already has some knowledge about. The users are instructed to query for additional resources, with which they can enrich their blog post while writing it.

The predetermined tasks alternate with tasks of the user's own choice, i.e. a possible sequence is (all tasks executed within the browser):

1. annotate a web page
2. read newspaper article
3. annotate a web page
4. write a blog entry
5. watch funny video clips
6. annotate a web page
7. engage in a forum discussion
8. annotate a web page
9. ...

This provides examples of tasks and session breaks as described in paragraph 4.2.2. Accordingly, the users are obliged to name the tasks of their own choice and indicate if recommendations are desirable for the particular task.

Software for Test Data Acquisition An adapted version of the browser extension prototype described in section 4.4 is used for conducting the user tests. To meet the requirements of the test data acquisition plan, the prototype is enriched with the following functionality: the possibility to record the performed tasks and related topics is added, code is injected into every web page to prevent interaction without having started a task and additional logging of user interactions is performed. Furthermore, recommendations are not provided automatically, but the user has to phrase the search query by herself. The changes are described in detail in section 4.4.4.

4.3 Information Source and Targeted Component

The available sources of information in the web browser setting for the first prototype are listed in table 5, along with applicable mining technologies and the intended use of the information parts.

Table 5: mining sources, technologies and purpose of use

source	applied technology	used for
search queries (+context)	disambiguation ("VISA" after "flight")	query reformulation →WP3
direct user feedback (thumb up/down)	learn to rank, collaborative filtering, content-based filtering	result re-ranking, result selection →WP3; feedback integration →WP4
browsing history - page content - navigational paths	 topic detection task detection	 query (re-)formulation →WP3 result re-ranking →WP3
bookmarks	topic detection	query (re-)formulation →WP3
browser profile (version, plugins, ...)	simple mapping	result presentation →WP2
rich social media feedback	tbd	result re-ranking →WP3; feedback integration →WP4

Note, that the column "used for" is only an estimation, since at the current state, we cannot tell for sure, which parts of information are actually usable (due to privacy constraints) and which parts contribute a recognizable impact to the recommendation quality. In this column, we also indicate the work packages, which we estimate to make use of the information.

"Direct user feedback" corresponds to "resource relations" in our long-term user profile and is not limited to explicit feedback, such as ratings but includes implicit observations, such as clicks and views of an item as well. Although bookmarking is included in this information source already (indicating an even higher relevance of a recommendation, than simply viewing it), bookmarks are listed in a separate row as well, since the set of already existing bookmarks in the browser can also serve to deduce a user's interest.

4.4 Web-based Prototype

Our first prototype for the web setting on a desktop machine is an extension to Google's Chrome browser, which provides recommendations based on the contents of the current web page or the contents of a selected paragraph within this web page. As backend for search queries, we use the Europeana API[18], which can be easily exchanged by the privacy-proxy - recommender chain later on. In addition, it enables the user to annotate passages of a web page, either with free text or recommended resources. This functionality provides feedback on the usage of resources at one hand and aligns towards a bookmarking scenario on the other hand. Being able to bookmark pages alongside with annotations may be an additional incentive for using the EEXCESS-framework.

[18] http://pro.europeana.eu/api

4.4.1 Architecture

The basic architecture is already described in deliverable D1.1, specifically in Section 3.2.1, and we will focus on the details of the most important components here (the content and background script are described within this section, while the injected user interface is presented in section 4.4.2).

Content Script Separate instances of the content script are injected into every web page. Since the content script has full access to the DOM-tree of the current page, it is responsible for mining the contents of the current page and forwarding them to the background script. In a very first simple approach, we used the top-3 terms of the current page as query terms to retrieve recommendations.

Beneath mining the provided contents of the given web page, the content script also monitors the user's interactions with the page. This is essential in particular for content creation scenarios. By keeping track of the user inputs, the extension can deduce the information need based on recently entered text and hence provide according recommendations. Moreover it allows us to establish search histories from search engines the user habitually uses. While search histories provide a great information source for personalization strategies as presented in section 4.2.1, we do not expect users to extensively utilize the search interface in our injected sidebar (and do not intend having users formulate a massive amount of queries, but instead providing interesting recommendations automatically). Thus, we have search histories available without any additional effort of the user.

The third task of the content script is the injection of the user interface into the current web page via an *iframe*-element. The use of an iframe is necessary, in order to avoid inheriting CSS-styles of the current page. The basic injected user interface, which simulates the behavior of a sidebar and presents recommendation results in a simple list is described in section 4.4.2. Note, that the use of an iframe-element allows the injection of other user-interfaces, which are totally decoupled from the basic one as well. The injected user interface communicates directly with the background script via message-passing, thus, solely this communication requires a common interface, respectively common methods in the user interface to be called by the background script for presenting search result for example.

In addition, the content script performs the integration of annotations into a web page (for illustration, see figure 4, the markers xx in this paragraph reference specific parts in the figure). It retrieves already existing annotations from the background script, highlights affected text passages 07 and displays the according contents on hover, along with the possibility to edit them. When marking textual parts, the user is provided with the ability to create an annotation for the selection, either by entering a free text comment 08 or semantically tagging 09a the selection with a resource from the recommendations list in the EEXCESS user interface 09b. Newly created or edited annotations are sent to the background script for storage.

Background Script Only a single instance of the background script exists for the whole extension. This script is responsible for the construction of long- and short-term profile, based on the contents retrieved by the content scripts and browsing history, communication with the backends and logging of retrieved results and corresponding interactions (e.g. if a result has been viewed, rated, etc).

The background script also serves as a mediator for communication between the user interface injected in a web page and the content script, since these two cannot communicate directly.

Since the API for the browsing history provided by the Chrome API is not sufficient for our needs, the background script features an own history implementation. This history contains not only a timestamp for the occurrence of a visit, but instead stores the active dwell time on the particular page. "Active" means, the page has the focus (switching to another application means switching the focus and thus, the visit ends). This does not imply, that the page actually has the user's focus, as she may be away from the computer, with the page still open. A conceivable extension to overcome this limitation is a timeout, ending the visit after a certain amount of time, no user interaction (mouse clicks, scroll events, ...) has occurred. The extension has the "background" permission, which means it becomes alive with the operating system (before Chrome is started) and keeps running after Chrome is shut

down. This allows for capturing the end times of visits, given by closing the browser, which could not be captured otherwise. In addition, the referring URL (if any), which is important for the creation of navigational paths is stored explicitly along with the visit, while the Chrome API requires additional effort to retrieve it. Nevertheless, the visits contain the identifiers of the corresponding original visit items in the Chrome API.

Data storage in the background script is accomplished by an indexed database[19] (indexedDB) with details given in section 4.4.3

4.4.2 Basic User Interface

Figure 4 shows the basic user interface, injected in every web page, simulating the behavior of a sidebar. The user can switch the injection on or off respectively by clicking the EEXCESS extension icon 01 next to the location bar. The search terms which triggered the presented recommendation

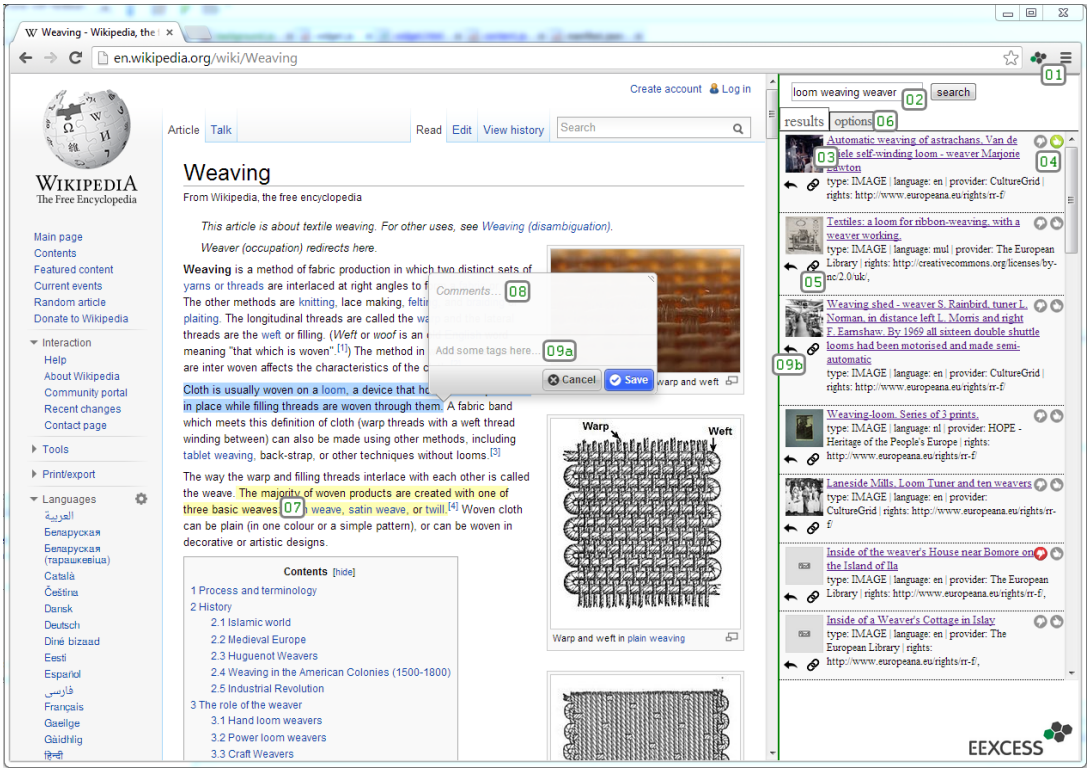

Figure 4: screenshot of injected user interface

results are displayed in the search field 02 and can be edited by the user. Adjustments to the query are logged, in order to be able to learn the user's query preferences. The result set list provides the title, a preview image, facets, such as type, language, provider and rights (if available) for each resource, along with additional interaction possibilities. By clicking either the title or the preview image, an overlay (see figure 5 for an example) is shown with an HTML representation of the resource, containing additional information. Presenting this information inside the current web page, provides us the ability to keep track of dwell time on a particular result directly within the extension. For each result, the user is provided with the ability to give explicit feedback, by either rating the result up or down with the buttons at 04. If a reference to the resource is required, e.g. in a content creation scenario, such as writing a blog entry, the URL of the resource can be retrieved by the button at 05.

[19] http://www.w3.org/TR/IndexedDB/

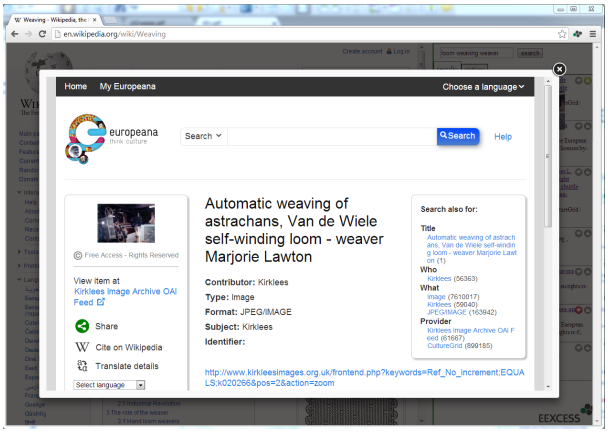

Figure 5: screenshot of overlay with additional information for a resource

The "options" tab 06 provides a link to the extension's options page, which is accessible via the extension administration page in Chrome as well. At the current state, it only provides the ability to enter and edit demographic information.

4.4.3 Data Storage

At the current point of time, all data handled within the prototype is stored locally on the client in an indexedDB. This database consists of object stores, holding records with key-value pairs. Valid keys are Numbers, Strings, Date-objects and Array-objects (with some limitations, e.g. the value of a number must not be "NaN"). The value can be any value supported by the structured clone algorithm[20], which provides some benefits over JSON-serialization, such as being able to duplicate Blob-, File- and FileList-objects for example, but is still not capable of function-serialization. Additional advantages of the indexedDB over Web Storage[21] (as alternative client-side storage possibility) are the possibility of storing duplicate values for keys and the efficient retrieval of records via indexes.

Annotations stored in the database conform to the Open Annotation Data Model[22] and are represented in the JSON-LD[23] format. Thus, they can be easily published and shared among different platforms in the future.

Nevertheless, storing the data solely on the client has one big disadvantage: when the user deletes his private data via the browser-integrated function, all of the extensions data are swept away. In order to provide continuously stable quality in personalized recommendations, at least some parts of the profile need to be stored on a server-side trusted third party (in case of the privacy-proxy being realized server-side, it could be a possible storage location).

4.4.4 Adaptions for Test Data Acquisition

This section describes adaptions to the prototype that were performed to meet the test setting requirements as presented in section 4.2.2. Basically, they comprise an extension of the injected widget's user interface to record the execution of tasks, functionality to prevent user interaction before having started a task, changes to the query process and logging of additional information. An update of the object stores' structure was necessary alongside with these modifications.

[20] http://www.w3.org/TR/html5/infrastructure.html#safe-passing-of-structured-data
[21] http://www.w3.org/TR/webstorage/
[22] http://www.openannotation.org/spec/core/
[23] http://json-ld.org/

UI controls for task detection For recording tasks, an additional tab was added to the injected widget's menu. Figure 6 shows the contents of this additional tab. The task to perform needs to be

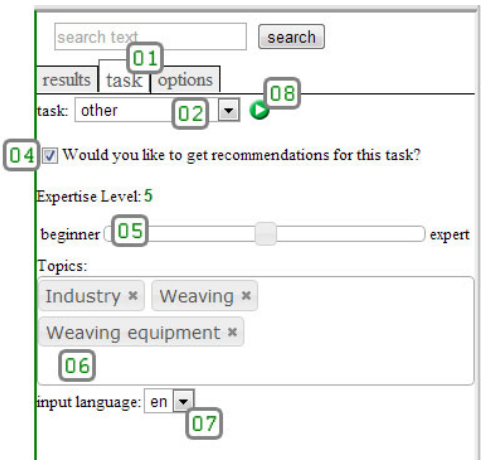

Figure 6: screenshot of task definition user interface

selected at the input field shown at 02. The user can select one of the predefined tasks ("annotate a webpage" or "write a blog entry") or choose "other". Choosing the latter will prompt to specify a custom label for the task after its execution. When choosing the task "other", an additional checkbox 04 is shown to indicate, whether recommendations are desirable for this task or not. The user can adjust his level of expertise on the task at hand with the slider at 05. Possible values range from 0 (lowest) to 10 (highest). The topics related to the specified task are defined at 06. This input field features auto-completion for dbpedia-categories. This means, the user gets suggested dbpedia-categories that possibly match her input while typing. 07 provides the selection of the language for which to execute the auto-completion (currently supported languages are French, English and German). After proper adjustment of all task settings, the user can start the task with the button 08. After a task has been started it is not possible to change its name anymore, while all other settings may still be adjusted, because they might not be known in advance. For instance, the topics related to a task are usually discovered during its execution and the level of expertise evolves accordingly.

4.5 Prototype for Mobile Devices

In this section we describe the conceptual idea of a first mobile application for user context detection for a just-in-time retrieval scenario [Rhodes, 2000]. The prototype is currently under active development and we report the current status here.

The scenario for this mobile context detection is a general public user interested in architecture visiting a city with cultural history. Depending on the location he or she should get recommendations which interesting places are nearby and what the current site had looked like during the history. But the user might not want suggestions if he is currently listening to music or writing an email. Thus more context information than just the physical location need to be aggregated for this scenario.

With this prototype we will investigate the following research questions:

- How can (physical) context in mobile devices be used to detect *what* to query from the database? In detail, this requires answers to the following questions
  - Where to get terms for the query (e.g. Notes, Messages, Calendar, Location)?
  - How to match the terms to the various fields of the query?
  - How to combine terms for queries?

- How can (physical) context be used in mobile devices to detect *when* to query from the database and present the results? More specific, this comprises the following questions:
  - In what situations will a user see the results as useful and not disturbing?
  - How often should the user be presented results?

In a first prototype for mobile phones we target the Android platform. We use the recently published AWARE framework[24]. AWARE is an open-source context-detection middleware for mobile devices running Android operation system. The framework provides access to a multitude of sensors, e.g., acceleration, gravity, gyroscope, location, bluetooth status.

As a backend database we use the Europeana API.

4.6 Summary

In this section we defined the EEXCESS user profile differentiating between long-term, short-term user profile and context. A literature review revealed possible directions for learning the user profile. Based on this review we will start with learning user interest first. Further, we described our plan for the acquisition of test data necessary for supervised approaches. The preparations are finalized, and the data acquisition is currently work in progress. Further, we presented two user mining prototypes, a Web-based browser plugin and the conceptual idea of a mobile prototype, which is currently under development.

[24] http://www.awareframework.com/

5 Privacy-Aspects

This section reviews work that has been done within WP5, tasks 5.2.

5.1 Relation to other Workpackages

In their review Micarelli et al. [Micarelli et al., 2007] show three principle ways to integrate user profile information into recommendations systems (see also figure 7):

- the user profile is tightly coupled to the retrieval process
- the user profile is used in a post-processing step, which basically means query re-ranking and/or filtering
- the user profile is used in a pre-processing step, which basically means query modification

Combined approaches are also possible, e.g. combining query-modification and result re-ranking (see for instance [Shen et al., 2005]).

Because EEXCESS's aims at privacy-preserving personalized recommendations, and no personal information may be transfered to the recommender, the user profile integration can only be implemented as pre- or post-processing step or a combination of both. The privacy policy and issues with respect to recommenders are discussed in detail in deliverable [D61, 2013].

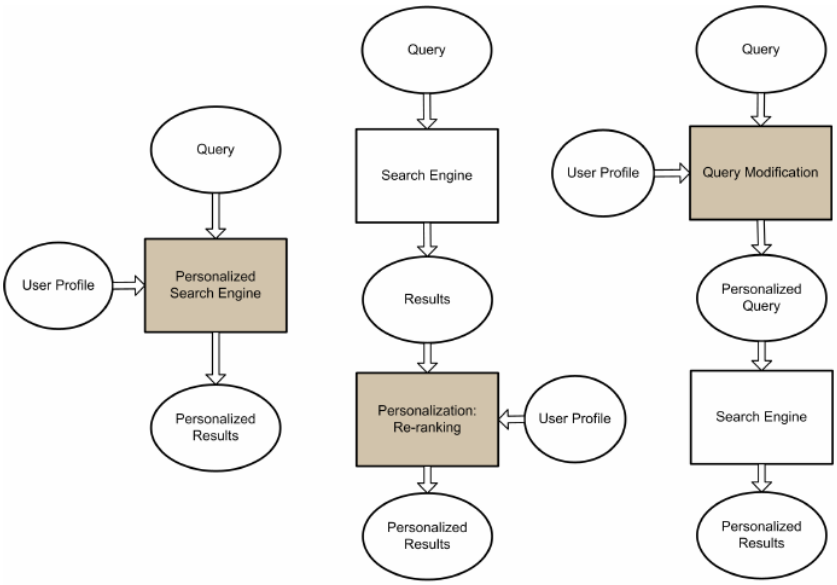

Figure 7: User model and recommender. User model is part of the retrieval process (left), post-processing (center), or pre-processing (right). Taken from [Micarelli et al., 2007]

The described principles apply to search scenarios, in which the query is explicitly specified by the user. Since EEXCESS aims to support the user in her currently executed task by recommending helpful resources automatically, such an explicit query is not necessarily given: it must be generated by the recommender itself. Thus, in order to create meaningful queries, the recommender requires at least some knowledge about a user's information need. In order to provide such information to the recommender without undermining privacy, the user profile must never be transferred to the recommender directly, but requires privacy conservation mechanisms in between, in particular the privacy preserving proxy. To bootstrap development of this transmission chain, we start not with

a full user model in the first year, but exchange only the user's weighted interests and contextual information with the privacy preserving proxy, which handles the anonymized propagation to the recommender.

5.2 Anonymizing User Profiles

Quite a number of technological approaches for recommender systems aiming at easing tension between personalization and users' privacy has been developed in recent years. One of them that has lately gained much attention is aggregation. It is based on the idea of maintaining users' privacy by hiding them in so-called interest groups, i. e. communities of users who share similar interests. For each of these anonymity sets, an aggregate group profile is computed and sent to the central server where recommendation is then done at a group level instead of at an individual level [Shang et al., 2013, Nandi et al., 2011]. That way, users' individual data is not disclosed to the service provider, but only some aggregated preference information.

A very critical step in this approach is the process of building such interest groups. On the one hand, in order not to reveal user's identity, those communities must not be too small. On the other hand, a rise in the size of the groups might result in a loss of recommendation quality as the aggregated profile becomes more inaccurate. Nevertheless, the question of how to group users respecting aforementioned constraints has not been studied yet. Therefore, the goal of this work will be to determine a way for setting up user interest groups assuring both privacy and a high recommendation quality.

In this context, the focus will be on the EEXCESS use case scenario of pushing valuable content to the user while he is editing a Wikipedia page in order to provide him with material for enriching this article with suitable additional content. The idea is to extract an interest profile from users' Wikipedia editing history. At this point, YAGO – a huge and clean ontology derived from Wikipedia and WordNet [Suchanek et al., 2007] – comes into play. With its help, a semantics-based hierarchical classification is done by mapping each entity – hence article the user is editing – onto a hierarchy of concepts describing the articles' topics on different granularity levels – from a very detailed description on the lowest level probably providing the most accurate recommendations to a more general and thus privacy-preserving one further up in the taxonomy (cf. figure 5.2).

The main problem now is to select an appropriate ontological depth for creating a user preference profile that both maintains users' privacy and provides high quality recommendations. Therefore, the following questions have to be answered:

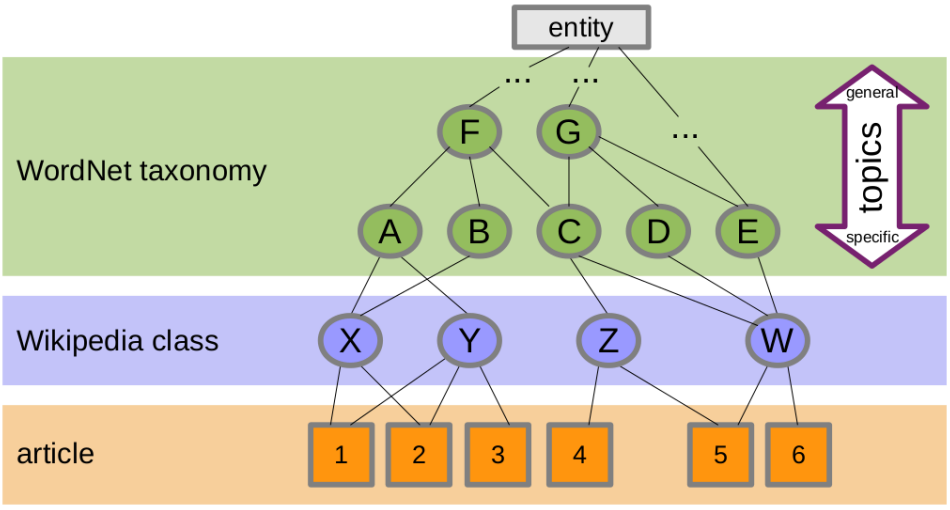

Figure 8: Settings for the privacy-experiment. Embedding of Wikipedia Articles in the YAGO ontology.

- On which hierarchical level can a user be described and still be hidden within an interest group?
- How specific should the topics describing user's interests be in order to yield accurate recommendations?

These two problems will be evaluated experimentally via a usage simulation based on statistics on Wikipedia page edits. As an entry point, an arbitrary Wikipedia article will be chosen. With the help of Wikipedia's "View history"-page, two editors of this article will then be selected. Next, by means of the "User contributions"-page the other Wikipedia articles they have edited will be identified. These pages reflect users' fields of interest and hence can be used to derive their preference information – thus, a user profile. With the help of the ontology YAGO it can then be determined on which granularity level both users can still be distinguished, and on which one this is not possible any more – implying that the user is hidden in an interest group and thus kept anonymous. Besides, it has to be evaluated if a user profile based on more specific topics indeed yields more accurate recommendations. This process has to be repeated dozens of times with a larger number of users in order to obtain statistically relevant results.

6 Usage Patterns of Resources

This section reviews work that has been done within WP5, specifically in tasks 5.3.

Is a memory, cultural, educational, or scientific organization (incl. its occupation area and services) noticed? Is it utilized? Who are the users? Is it valued, how do its users benefit from it? Does it reach a local, national, or international audience? Should marketing, presentation, or services be adapted? For organizations to get a comprehensive view on external comments in this regard, social media are to be included in the bunch of observed conversation channels. This task is motivated by such questions and aims at providing insight into the trends of targeted topics and audiences in EEXCESS relevant areas within social media channels.

6.1 Monitoring Social Media Channels

In the first phase we focus on tweets from Twitter[25], due to Twitter being popular, highly used (especially in the field of scholarly communications), and accessible via APIs. Before analyzing the communication stream for the EEXCESS partners, we start with considering topics and academics from the field of economics (ZBW content and focus group). Also, we start with a simple monitoring of tweets.

Twitter Characteristics Concrete, with a bottom-up view, information potentially conveyed by Twitter's single tweets comprises e.g. the actual statement, its language, topics, links and references, the author, other users (e.g. in case of comments or retweets), a geolocation, subscribers of the author/sender (the so-called followers), and an implicit sentiment. So-called hashtags, like e.g. #GenderPayGap, may explicitly state the topics of tweets. However, the #tags are freely assigned by the authors of the tweets. Graphs of tweets can exhibit relations between tweets based on some data field, as e.g. subscriptions (followee-follower), interactions (tweets-retweets/replies, revealing effective audiences of users or tweets), and topics (#tags, revealing discussion threads).

Basic Economic Vocabulary Statistics ZBW maintains a thesaurus for economics (STW[26]), which assembles a broad coverage of the concepts and phrases used in scholarly communication in the field of economics. This thesaurus contains about 5,800 concepts with about 32,000 describing terms. There also exist cross-concordances to other thesauri or taxonomies which are published as Linked Open Data (LOD). This increases the amount of available terms to about 37,000. Beside specifically economics terms and concepts, the STW also contains quite general terms and concepts (e.g. "government", "artist", "theory", "USA"). Just to get a very rough overview on the usage of economics vocabulary in Twitter, the Twitter REST API was queried with STW terms:

- Around 59% of the terms appeared at least once.
- About 60% of the terms appeared up less than ten times.
- On the other hand 94% of the concepts were referenced at least once.
- The median usage was about 130 times (since we only counted the first 100 matches, the median might have been be larger). In total, terms from the thesaurus were used more than one million times.

While absolute observations already yield some findings, relative studies like distributions, evolutions, and comparisons seem much more interesting. Such relative views may be exemplified by considering the traffic on similar issues, the traffic before and after some event, or the general traffic or spirit in a specific media channel.

[25] https://twitter.com/
[26] http://zbw.eu/stw/versions/latest/about.en.html

6.1.1 Basal Limitations on Completeness and Accuracy

Completeness and accuracy of the results depend on general circumstances, specifics of the Twitter API, and the chosen approach. On the other hand, the approach is to be aligned along issues of completeness and accuracy. General limitations are the following:

- Because the composition of the target audience is not completely known and dynamic, its tweets cannot be tracked exhaustively. Monitoring and archiving the complete Twitter stream is out of the scope of this project.
- Since members of the target audience may well also twitter on issues having nothing soever to do with the fields of the EEXCESS partners, supposedly not all of their tweets are relevant.
- Due to the terms in tweets not complying with official vocabularies (folksonomies), relevant tweets cannot be tracked exhaustively.
- Because words tend to be ambiguous, and anyway assigning #tags is up to the users, the actual topics of tweets are not faultlessly revealed by #tags.

The Twitter APIs also are limited in the following ways:

- Naturally, only publicly visible tweets can be taken into account, restricted tweets cannot.
- Provision of tweets by Twitter is not complete or not abiding/durable.
- Search in the REST API is limited to 180 queries per 15 minutes per access token[27].
- The REST search API "is not meant to be an exhaustive source of Tweets. Not all Tweets will be indexed or made available via the search interface." It provides "recent or popular" tweets.[28]
- The streaming API limits the number of tweets to "a small fraction of the total volume of Tweets"[29].
- The streaming API allows to filter with up to 400 cues (so-called track keywords) and 5,000 user IDs (follow userids)[30].
- Twitter also sets limits that are not made public[31].

6.1.2 Approach

Two issues have to be solved. First, one needs to collect relevant posts. Second, these posts have to be analyzed and presented in a beneficial manner.

Collecting Relevant Posts There are two ways to collect posts relevant to EEXCESS via the Twitter APIs, which might provide complementary information:

- Track users belonging to the target audiences.
- Track topics/keywords associated with the occupation area of EEXCESS partners.

[27] https://dev.twitter.com/docs/rate-limiting/1.1
[28] https://dev.twitter.com/docs/api/1.1, https://dev.twitter.com/docs/using-search
[29] https://dev.twitter.com/docs/faq#6861
[30] https://dev.twitter.com/docs/api/1.1/post/statuses/filter
[31] e.g. https://dev.twitter.com/docs/rate-limiting/1#search

When collecting tweets, the main goal is to catch as many relevant tweets as possible (high recall), especially so to deal with the limitations of the Twitter APIs, and also to capture as few irrelevant tweets as feasible (high precision).

To get a first impression of the general data flow and to get an estimate of the data volume and rate, we start with tracking a static list of users (i.e. twitter accounts) related to economics. We expect that tweets crawled from topic-related users will have a higher precision than tweets crawled with respect to keywords which might either be broadly used in other contexts or be so specific that the probability of use in twitter is low.

We compiled the list of users from two sources:

- A list of economic academics highly active on Twitter, curated by ZBW.
- The directory Wefollow[32], listing users with interest in certain topics like e.g. 'economics' or 'finance'. Wefollow is curated by the users themselves and incentively provides them so-called prominence scores. It was used as gold standard by[Pennacchiotti and Popescu, 2011].

The final list hits the limit of Twitter's streaming API of 5,000 users. For those users, we currently observe a rate of about 80,000 tweets per week. Additionally, we crawled tweets from the past for these users via the REST API.

This approach can be improved in several ways to overcome obvious limitations.

To increase recall we will therefore at first also use relevant keywords to collect tweets. Initially this might be a static list of keywords taken from the STW. The crawling criteria will then be made dynamic, i.e. we will adapt the set of users as well as the set of keywords over time. E.g. to adapt keywords [Li et al., 2013] proposed an algorithm to select a set of keywords which optimizes recall. They report a high recall even for only 20 keywords. Although, it must be investigated if this also applies to the vocabulary used by our target audience. The set of users might be extended or adjusted e.g. by users connected to relevant resources or by tracing followee-follower or interaction relations.

To increase precision a common approach is to filter the crawled tweets with the help of a classifier, which had been trained from manually labeled tweets. Two issues have to be considered with this approach. First, it must be evaluated if compiling training data for each partner is feasible. Second, the classifier must also be adapted to the used vocabulary over time. Lin et al. [Lin et al., 2011] studied the task of filtering a stream of tweets for tweets related to implicit topics of 'persistent' hashtags, i.e. rather broad hashtags 'that are (relatively) stable over time' (e.g. #travel or #health). The focus on topics which are implicitly labelled by hashtags enables the authors to train a classifier without explicitly labeling training data. Further, those implicit labels in the stream allow them to adapt the classifier over time due to their persistent nature. In our task we do not have implicit labels like persistent hashtags, but we may identify indicative features of tweets like if a tweet contains a link to a scholarly article, twitter account, or event in the realm of the EEXCESS partners. We will therefore investigate solutions based on similar ideas.

Analyzing Posts and Providing Results At first, we will provide simple statistics over the global dataset of tweets, e.g. on most used hashtags, most active users, histogram of interaction, overall number of tweets with relevant topics, etc. To rank users along their influence, various measures have been proposed (e.g, [Cha et al., 2010, Bakshy et al., 2011]).

For a finer grained focus of interest, more detailed insight can be gained on the level of entities like users or topics. A promising approach to analyze the evolution of a topic was demonstrated by [Fisher et al., 2008]. They use a timeline graph to visualize the number of blog entries referring to a news article. The peaks of the graph get annotated with the most frequent keywords at the corresponding time. This allows a user to get a grasp of how the focus of the discussion in the blogosphere changed over time. This approach can be adopted to visualize the evolution of discussions in twitter surrounding a specific topic.

[32] http://wefollow.com

Further, more specific needs of the EEXCESS partners will be evaluated when the implementation progresses.

Implementation The first implementation will be done as a proof of concept. Second to an initial working version, a more generic solution which is decoupled from the EconBiz use case is intended.

The current approach based on the static list of users has so far collected about two million tweets at a rate of roughly 80,000 tweets per week. The tweets are indexed in Elasticsearch[33]. For simple analyses Kibana[34] is used as a frontend.

6.2 Summary

In this section we described the first approach to mine usage data of EEXCESS resources with focus on social media channels. Based on a first analysis on Twitter streams we identified the next steps as collecting basic statistics of related users and resources over a longer period of time.

[33] http://www.elasticsearch.org/
[34] http://www.elasticsearch.org/overview/kibana/

7 Summary and Future Work

In this deliverable we presented the EEXCESS user profile and first approaches automatic population. From the resource perspective we identified different usage mining components and an conceptual approach idea for resource usage mining in Twitter. We also presented the route we will research for user interest anonymization. Because the privacy aspects are the focus of WP 6 we refer to [D61, 2013] for a broader overview. The main questions we identified with respect to the user profile are the following (see an overview in figure 9):

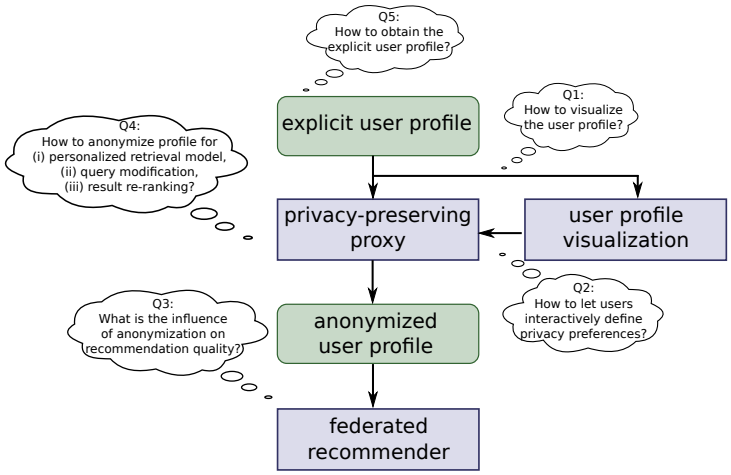

Figure 9: Research questions related to the user profile in EEXCESS

Q1: How to visualize the user profile? In the requirements specification the need for users to be able to view and adapt their profiles was identified (see D1.1., Section 4.4.). Because the EEXCESS user profile is rather complex (see Section 4.1, we aim to use Information Visualization approaches to make it accessible and modifiable by users. We think that researching towards adapting open learner models [Bull and Kay, 2012] is a promising way.

Q2: How can users interactively define privacy preferences? Depending on the privacy policy (outcome of WP 6), the user must be able to define privacy settings herself. To define (new) settings, the user must be aware of current settings, and optimally of potential influences on the recommendation accuracy. This question is related to Q1, where the privacy settings can be visualized with the user profile.

Q3: What is the influence of anonymization to the quality of recommendations? Due to privacy considerations the user profile in EEXCESS needs to be anonymized. How different levels of anonymization influence the recommendation accuracy is an open issue.

Q4: How should the user profile be anonymized for the 3 types of personalization? Depending on the final recommendation strategy within EEXCESS, the anonymization strategy of user profiles may differ (see section 5.1 for the different strategies. This research question is in the focus of WP6.

Q5: How to obtain the explicit user profile? The most central research question for WP 5 is how the explicit user profile is populated. This question spawns multiple new research questions, refering to the different parts of the user profile. E.g, “how can we obtain topics of interest for a user based on her browsing behavior”.

The next step will be to research the learning of the user profile (Question 5). In the first prototype we will perform the learning of the user profile in an offline training stage with the data acquired as described in section 4.2.2 Depending on privacy considerations of WP 6 and the final location of the proxy (client or server) we will then either

- research embedded machine learning approaches to do the learning solely on client-side, possibly exchanging trained machine learning models to the client. These models would be trained on data from test users which have no privacy concerns.
- perform server-side machine learning and exchange the user profile from server to client
- research on learning methods for specific parts of the user profile which have been identified as privacy-sensitive

User Profile for First Integrated Prototype In the first integrated prototype the automatic detection mechanisms will not be fully researched yet and also their integration into the federated recommender will not be fully solved. Regarding the long-term profile, our focus in the first approach is on the features, for which we expect the highest impact (i.e. interests and tasks), as stated in section 4.1.2. Partly, we will cover resource relations as well, in particular the fraction that can be addressed by direct feedback in the client application and thus can be mined rather easily. Although we expect a high learning complexity for tasks, we address them right from the beginning, since they seem the key factor to determine when to provide recommendations. The only additional feature to interest and task in the short-term profile is the session, which is highly correlated to tasks and thus not excluded from first targeted features. We reduce the set of contextual features to be evaluated in the first year to geo-location and the user's most narrow focus (i.e. the currently viewed, or - even narrower - selected paragraph in a web page).

The user profile to be transferred among the system is further reduced to the set of weighted interests and the two contextual features just mentioned. Since this model comprises personal information, it must not be accessible by components outside the client, except for the privacy preserving proxy, which will take care of its further processing, described in deliverable [D61, 2013].

Usage Mining for First Integrated Prototype In the first integrated prototype there will be practically no privacy concerns, meaning that the user and usage mining component is a trusted component. Therefore we will collect all necessary information on the client and transfer them to a server responsible for logging. This logging is solely for collection of measures necessary to obtain the project specific success factors.

7.1 Future Work

Apart from the work on populating the user profile, proof-of-concept usage mining and profile anonymization we identified three main directions for future work.

First, currently the EEXCESS system is a closed system with respect to user mining. This means all information we have for users will be gathered within the project components. A future direction will be to also collect user information from external sources, such as social networks. This requires further research on user profile integration.

Second, the usage mining components are stand-alone proof-of-concept components. In the future we need to investigate how the collected information is integrated to a full usage picture.

Third, we need to extend the current definition of the user profile to also include privacy-settings of users. This means the user should be able to rate the sensitivity of her profile information. WP 5 will research the modeling aspects, whereas in WP 2 the user interface aspects are covered and WP 6 will research on how to integrate this information into the anonymization process.

8 Glossary

BITM

BitMedia, Austria

CT

Collection Trust, United Kingdom

CUR

Client-side User and Resource mining

DoW

Description of Work

EC

European Commission

EEXCESS

Enhancing Europe's eXchange in Cultural Educational and Scientific Resources

ESR

EEXCESS Server-side Resource mining

INSA

Institut National des Sciences Appliquées (INSA) de Lyon, France

JR-DIG

JOANNEUM RESEARCH Forschungsgesellschaft mbH, Austria

KBL-AMBL

Kanton Basel Land, Suisse

Know-Center

Kompetenzzentrum für Wissenschaftsbasierte Anwendungen und Systeme Forschungs- und Entwicklungs Center GmbH, Austria

MEN

Mendeley Ltd., United Kingdom

PPM

Privacy-Procy mining

PSR

Partner Server-side Resource mining

Uni Passau

University of Passau, Germany

WM

wissenmedia, Germany

ZBW

German National Library of Economics, Germany

Acknowledgement

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 600601.

9 References

[D11, 2013] (2013). D1.1 – First Conceptual Architecture and Requirements Definition. Technical report.

[D41, 2013] (2013). D4.1 – Integration and Enrichment Specification Analysis. Technical report.

[D61, 2013] (2013). D6.1 – Policy Model for Privacy Preservation and Feasibility Report. Technical report.

[Bakshy et al., 2011] Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Everyone's an influencer. In *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11*, page 65, New York, New York, USA. ACM Press.

[Bazire and Brézillon, 2005] Bazire, M. and Brézillon, P. (2005). Understanding context before using it. In *Proceedings of the 5th international conference on Modeling and Using Context*, CONTEXT'05, pages 29–40, Berlin, Heidelberg. Springer-Verlag.

[Bennett et al., 2012] Bennett, P. N., White, R. W., Chu, W., Dumais, S. T., Bailey, P., Borisyuk, F., and Cui, X. (2012). Modeling the impact of short- and long-term behavior on search personalization. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 185–194, New York, NY, USA. ACM.

[Bilenko and Richardson, 2011] Bilenko, M. and Richardson, M. (2011). Predictive client-side profiles for personalized advertising. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 413–421, New York, NY, USA. ACM.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

[Brusilovsky and Millán, 2007] Brusilovsky, P. and Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In Brusilovsky, P., Kobsa, A., and Nejdl, W., editors, *The adaptive web*, pages 3–53. Springer-Verlag, Berlin, Heidelberg.

[Bull and Kay, 2012] Bull, S. and Kay, J. (2012). *International Handbook of Metacognition and Learning Technologies*, chapter Open Learner Models. Springer.

[Buzikashvili, 2006] Buzikashvili, N. (2006). Automatic task detection in the web logs and analysis of multitasking. In Sugimoto, S., Hunter, J., Rauber, A., and Morishima, A., editors, *Digital Libraries: Achievements, Challenges and Opportunities*, volume 4312 of *Lecture Notes in Computer Science*, pages 131–140. Springer Berlin Heidelberg.

[Calegari and Pasi, 2013] Calegari, S. and Pasi, G. (2013). Personal ontologies: Generation of user profiles based on the {YAGO} ontology. *Information Processing & Management*, 49(3):640 – 658. <ce:title>Personalization and Recommendation in Information Access</ce:title>.

[Carberry et al., 2013] Carberry, S., Weibelzahl, S., Micarelli, A., and Semeraro, G., editors (2013). *User Modeling, Adaptation, and Personalization - 21th International Conference, UMAP 2013, Rome, Italy, June 10-14, 2013, Proceedings*, volume 7899 of *Lecture Notes in Computer Science*. Springer.

[Cha et al., 2010] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy.

[Chemudugunta et al., 2008] Chemudugunta, C., Holloway, A., Smyth, P., and Steyvers, M. (2008). Modeling documents by combining semantic concepts with unsupervised statistical learning. In *International Semantic Web Conference*, Lecture Notes in Computer Science, pages 229–244. Springer.

[Claypool et al., 2001] Claypool, M., Le, P., Wased, M., and Brown, D. (2001). Implicit interest indicators. In *Proceedings of the 6th International Conference on Intelligent User Interfaces*, IUI '01, pages 33–40, New York, NY, USA. ACM.

[Daoud et al., 2009] Daoud, M., Tamine-Lechani, L., Boughanem, M., and Chebaro, B. (2009). A session based personalized search using an ontological user profile. In *Proceedings of the 2009 ACM symposium on Applied Computing*, pages 1732–1736. ACM.

[Dey et al., 2001] Dey, A. K., Abowd, G. D., and Salber, D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Hum.-Comput. Interact.*, 16(2):97–166.

[Dou et al., 2007] Dou, Z., Song, R., and Wen, J.-R. (2007). A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th international conference on World Wide Web*, pages 581–590. ACM.

[Dourish, 2004] Dourish, P. (2004). What we talk about when we talk about context. *Personal Ubiquitous Comput.*, 8(1):19–30.

[Fischer, 2001] Fischer, G. (2001). User modeling in human-computer interaction. *User Modeling and User-Adapted Interaction*, 11(1-2):65–86.

[Fisher et al., 2008] Fisher, D., Hoff, A., Robertson, G., and Hurst, M. (2008). Narratives: A visualization to track narrative events as they develop. In *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on*, pages 115–122. IEEE.

[Fox et al., 2005] Fox, S., Karnawat, K., Mydland, M., Dumais, S., and White, T. (2005). Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.*, 23(2):147–168.

[Granitzer et al., 2009] Granitzer, M., Rath, A. S., Kröll, M., Seifert, C., Ipsmiller, D., Devaurs, D., Weber, N., and Lindstaedt, S. (2009). Machine learning based work task classification. *Journal of Digital Information Management (JDIM)*, 7(5).

[Heckmann et al., 2005] Heckmann, D., Schwartz, T., Brandherm, B., Schmitz, M., and von Wilamowitz-Moellendorff, M. (2005). Gumo - the general user model ontology. In *User Modeling*, pages 428–432.

[Joachims et al., 2005] Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 154–161, New York, NY, USA. ACM.

[Kellar et al., 2004] Kellar, M., Watters, C., Duffy, J., and Shepherd, M. (2004). Effect of task on time spent reading as an implicit measure of interest. *Proceedings of the American Society for Information Science and Technology*, 41(1):168–175.

[Kelly and Belkin, 2004] Kelly, D. and Belkin, N. J. (2004). Display time as implicit feedback: Understanding task effects. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 377–384, New York, NY, USA. ACM.

[Kosinski et al., 2013] Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*.

[Li et al., 2007] Li, L., Yang, Z., Wang, B., and Kitsuregawa, M. (2007). Dynamic adaptation strategies for long-term and short-term user profile to personalize search. In Dong, G., Lin, X., Wang, W., Yang, Y., and Yu, J., editors, *Advances in Data and Web Management*, volume 4505 of *Lecture Notes in Computer Science*, pages 228–240. Springer Berlin Heidelberg.

[Li et al., 2013] Li, R., Wang, S., and Chang, K. C.-C. (2013). Towards social data platform: automatic topic-focused monitor for twitter stream. *Proceedings of the VLDB Endowment*, 6(14):1966–1977.

[Lin et al., 2011] Lin, J., Snow, R., and Morgan, W. (2011). Smoothing techniques for adaptive online language models. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*, page 422, New York, New York, USA. ACM Press.

[Liu et al., 2005] Liu, T.-Y., Yang, Y., Wan, H., Zeng, H.-J., Chen, Z., and Ma, W.-Y. (2005). Support vector machines classification with a very large-scale taxonomy. *SIGKDD Explor. Newsl.*, 7(1):36–43.

[Matsuo, 2003] Matsuo, Y. (2003). Word weighting based on user's browsing history. In Brusilovsky, P., Corbett, A., and Rosis, F., editors, *User Modeling 2003*, volume 2702 of *Lecture Notes in Computer Science*, pages 35–44. Springer Berlin Heidelberg.

[Matthijs and Radlinski, 2011] Matthijs, N. and Radlinski, F. (2011). Personalizing web search using long term browsing history. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 25–34. ACM.

[Micarelli et al., 2007] Micarelli, A., Gasparetti, F., Sciarrone, F., and Gauch, S. (2007). Personalized search on the world wide web. In Brusilovsky, P., Kobsa, A., and Nejdl, W., editors, *The adaptive web*, chapter Personalized search on the world wide web, pages 195–230. Springer-Verlag, Berlin, Heidelberg.

[Morita and Shinoda, 1994] Morita, M. and Shinoda, Y. (1994). Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 272–281, New York, NY, USA. Springer-Verlag New York, Inc.

[Murray et al., 2007] Murray, G. C., Lin, J., and Chowdhury, A. (2007). Identification of user sessions with hierarchical agglomerative clustering. *Proceedings of the American Society for Information Science and Technology*, 43:1–9.

[Nandi et al., 2011] Nandi, A., Aghasaryan, A., and Bouzid, M. (2011). P3: A privacy preserving personalization middleware for recommendation-based services. In *Hot Topics in Privacy Enhancing Technologies Symposium*.

[Pennacchiotti and Popescu, 2011] Pennacchiotti, M. and Popescu, A.-M. (2011). Democrats, republicans and starbucks afficionados: User classification in twitter. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 430–438, New York, NY, USA. ACM.

[Poese et al., 2011] Poese, I., Uhlig, S., Kaafar, M. A., Donnet, B., and Gueye, B. (2011). Ip geolocation databases: Unreliable? *SIGCOMM Comput. Commun. Rev.*, 41(2):53–56.

[Rath et al., 2008] Rath, A. S., Kröll, M., Lindstaedt, S., Weber, N., Granitzer, M., and Dietzel, O. (2008). Context-aware Knowledge Services. In *Proceedings of Computer Human Interaction (CHI 2008), Workshop on Personal Information Management: PIM 2008*.

[Rhodes, 2000] Rhodes, B. J. (2000). *Just-In-Time Information Retrieval*. PhD thesis, Massachusetts Institute of Technology.

[Rubin et al., 2012] Rubin, T. N., Chambers, A., Smyth, P., and Steyvers, M. (2012). Statistical topic models for multi-label document classification. *Mach. Learn.*, 88(1-2):157–208.

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

[Shang et al., 2013] Shang, S., Hui, Y., Hui, P., Cuff, P. W., and Kulkarni, S. R. (2013). Privacy preserving recommendation system based on groups. *CoRR*, abs/1305.0540.

[Shen et al., 2005] Shen, X., Tan, B., and Zhai, C. (2005). Implicit user modeling for personalized search. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 824–831, New York, NY, USA. ACM.

[Sontag et al., 2012] Sontag, D., Collins-Thompson, K., Bennett, P. N., White, R. W., Dumais, S., and Billerbeck, B. (2012). Probabilistic models for personalizing web search. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 433–442. ACM.

[Suchanek et al., 2007] Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). YAGO: a core of semantic knowledge unifying WordNet and Wikipedia. In *WWW '07: Proceedings of the 16th International World Wide Web Conference, Banff, Canada*, pages 697–706.

[Teevan et al., 2008] Teevan, J., Dumais, S. T., and Liebling, D. J. (2008). To personalize or not to personalize: modeling queries with variation in user intent. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 163–170. ACM.

[Turnina, 2013] Turnina, A. (2013). A novel approach for modeling users short-term interests, based on user queries. *International Journal of Computer Science Issues*, 10(1).

[Zandbergen, 2009] Zandbergen, P. A. (2009). Accuracy of iphone locations: A comparison of assisted gps, wifi and cellular positioning. *Transactions in GIS*, 13:5–25.

A Appendix

A.1 User Profile Example "Horst B."

```
<?xml version="1.0"?>
<!DOCTYPE rdf:RDF [
<!ENTITY owl "http://www.w3.org/2002/07/owl#">
<!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#">
<!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#">
<!ENTITY xsd "http://www.w3.org/2001/XMLSchema#">
<!ENTITY dc "http://purl.org/dc/elements/1.1/">
<!ENTITY dcterms "http://purl.org/dc/terms/">
<!ENTITY gumo "http://gumo.org/2.0/">
<!ENTITY ubis "http://ubisworld.org/documents/ubis.rdf#">
<!ENTITY ubisworld "http://ubisworld.org/show.php?">
<!ENTITY foaf "http://xmlns.com/foaf/0.1">
]>
<rdf:RDF
        xmlns:owl = "&owl;"
        xmlns:rdfs = "&rdfs;"
        xmlns:rdf = "&rdf;"
        xmlns:xsd = "&xsd;"
        xmlns:dc = "&dc;"
        xmlns:dcterms = "&dcterms;"
        xmlns:gumo = "&gumo;#"
        xmlns:ubis = "&ubis;"
        xmlns:foaf = "&foaf;"
        xmlns:wi = "http://purl.org/ontology/wi/core#"
        xmlns:wo = "http://purl.org/ontology/wo/core#"
        xmlns:tl = "http://purl.org/NET/c4dm/timeline.owl#"
        xmlns:oa = "http://www.w3.org/ns/oa#"
        xmlns:rev = "http://purl.org/stuff/rev#"
        xmlns:tmo = "http://www.semanticdesktop.org/ontologies/2008/05/20/tmo#"
        xmlns:base ="http://www.eexcess.eu/up#" >

<foaf:Person rdf:ID="HorstB">
<!-- ########################################## DEMOGRAPHICS ################################################### -->
        <foaf:knows rdf:resource="http://www.bitmedia.cc/EntrepreneursSkillsCertificate"/>
        <foaf:name>Horst B</foaf:name>
        <foaf:title>Mag.</foaf:title>
        <foaf:givenname>Horst</foaf:givenname>
        <foaf:family_name>B</foaf:family_name>
        <foaf:workplaceHomepage rdf:resource="http://www.gymtamsweg.at/"/>

        <gumo:birthday>01.02.1973</gumo:birthday>
        <!-- no required format specified, expected impact very low, thus represented by a string -->
        <gumo:birthplace>Tamsweg, Austria</gumo:birthplace>
        <!-- Format needs to be defined, currently only String -->
        <gumo:educationlevel>University/master study</gumo:educationlevel>
        <gumo:city>Tamsweg</gumo:city>
        <gumo:street>Lasabergweg</gumo:street>
        <gumo:housenumber>12</gumo:housenumber>
        <gumo:postalcode>5580</gumo:postalcode>
        <gumo:country>Austria</gumo:country>

<!-- ########################################## PREFERENCES ############################################################ -->
        <!-- ### INTERESTS ### -->
        <wi:preference>
                <wi:WeightedInterest>
                        <wi:topic rdf:resource="http://dbpedia.org/data/Category:Running"/>
                        <wi:overall_weight>
                                <wo:Weight>
                                        <wo:scale rdf:resource="0to10scale" />
                                        <wo:weight_value rdf:datatype="xsd:decimal">3.0</wo:weight_value>
                                </wo:Weight>
                        </wi:overall_weight>
                </wi:WeightedInterest>
        </wi:preference>

        <wi:preference>
                <wi:WeightedInterest>
                        <wi:topic rdf:resource="http://dbpedia.org/data/Category:Football" />
                        <wi:overall_weight>
                                <wo:Weight wo:weight_value="8.0">
                                        <wo:scale rdf:resource="0to10scale" />
                                </wo:Weight>
                        </wi:overall_weight>
                </wi:WeightedInterest>
        </wi:preference>

        <wi:preference>
                <wi:WeightedInterest>
                        <wi:topic rdf:resource="http://dbpedia.org/data/Category:Skiing" />
                        <!-- Skiing just in winter -->
                        <wi:interestDynamics>
                                <wi:InterestDynamics rdf:about="winter">
                                        <wi:appear_time>
                                                <tl:Interval>
                                                        <tl:start rdf:datatype="xsd:gMonth">--12</tl:start>
                                                        <tl:duration>P3M</tl:duration>
                                                </tl:Interval>
                                        </wi:appear_time>
                                        <wo:weight>
                                                <wo:Weight>
                                                        <wo:scale rdf:resource="0to10scale"/>
                                                        <wo:weight_value rdf:datatype="xsd:decimal">7.0</wo:weight_value>
                                                </wo:Weight>
                                        </wo:weight>
                                </wi:InterestDynamics>
```

```
                        </wi:interestDynamics>
                </wi:WeightedInterest>
        </wi:preference>

        <!-- ### KNOWLEDGE ### -->
        <wi:preference>
                <base:WeightedKnowledge>
                        <wi:topic rdf:resource="http://dbpedia.org/data/Category:Payment_systems" />
                        <wi:overall_weight>
                                <wo:Weight wo:weight_value="9.0">
                                        <wo:scale rdf:resource="0to10scale" />
                                </wo:Weight>
                        </wi:overall_weight>
                </base:WeightedKnowledge>
        </wi:preference>

<!-- ################################### PUBLICATIONS (resource relation) ######################################### -->
        <foaf:publication rdf:resource="#articleHS00"/>
</foaf:Person>

<!-- ################################################ TASKS ############################################################ -->
<tmo:Task>
        <tmo:taskName>Searching for Literature</tmo:taskName>
        <tmo:subTask rdf:resource="#findConferences" />
        <tmo:subTask rdf:resource="#searchACM" />
        <base:frequency>
                <wo:Weight>
                        <wo:scale rdf:resource="0to10scale"/>
                        <wo:weight_value rdf:datatype="xsd:decimal">7.0</wo:weight_value>
                </wo:Weight>
        </base:frequency>
</tmo:Task>

<tmo:Task rdf:ID="findConferences">
        <tmo:taskName>Find Conferences</tmo:taskName>
</tmo:Task>
<tmo:Task rdf:ID="searchACM">
        <tmo:taskName>Search ACM Digital Library Catalog</tmo:taskName>
</tmo:Task>

<!-- ###################################### SOCIAL CONNECTIONS ################################################### -->
<foaf:Group rdf:about="Klasse2006B">
        <foaf:name>Klasse 2006 B</foaf:name>
        <rdfs:comment xml:lang="en">corresponds to the school class B of pupils who entered the school in 2006 </rdfs:comment>
        <foaf:homepage rdf:resource="http://www.gymtamsweg.at/2006B"/>
        <foaf:member rdf:resource="#HorstB" />
</foaf:Group>

<!-- ################################### RESOURCES AND RELATIONS ############################################## -->
<oa:Annotation>
        <oa:hasTarget>http://www.europeana.eu/portal/record/2022343/3C1926A37DB1C516C17FF8511A3CFF8C79D390BB.html</oa:hasTarget>
        <oa:annotatedBy rdf:resource="#HorstB" />
        <oa:hasBody>
                <oa:SemanticTag>
                        <!-- provide general rating class (including max/min)? | use wo? (with own class?) -->
                        <rev:rating>3.0</rev:rating>
                                <rev:minRating>1.0</rev:minRating>
                                <rev:maxRating>5.0</rev:maxRating>
                </oa:SemanticTag>
        </oa:hasBody>
</oa:Annotation>

<foaf:Document rdf:ID="articleHS00">
        <foaf:topic rdf:resource="http://dbpedia.org/resource/Zahlungssysteme"/>
        <dc:title>Historische Entwicklung von Zahlungssystemen</dc:title>
        <rdfs:comment xml:lang="en">this is an article written by Horst S.</rdfs:comment>
</foaf:Document>

<!-- ################################################ MISC ############################################################ -->
<wo:Scale rdf:about="0to10scale">
        <wo:min_weight>0.0</wo:min_weight>
        <wo:max_weight>10.0</wo:max_weight>
        <wo:step_size>1.0</wo:step_size>
</wo:Scale>

<rdfs:Class rdf:about="base:WeightedKnowledge">
        <rdfs:subClassOf rdf:resource="wi:WeightedInterest" />
</rdfs:Class>

<rdf:Property rdf:about="base:frequency">
        <rdfs:label>hasFrequency</rdfs:label>
        <rdfs:comment>The frequency of a task is represented by a corresponding weight. Thus, it depicts, if it is a commonly executed task,
               or if the task occurs rarely</rdfs:comment>
        <rdfs:domain rdf:resource="tmo:Task"/>
        <rdfs:range rdf:resource="wo:Weight"/>
</rdf:Property>
</rdf:RDF>
```

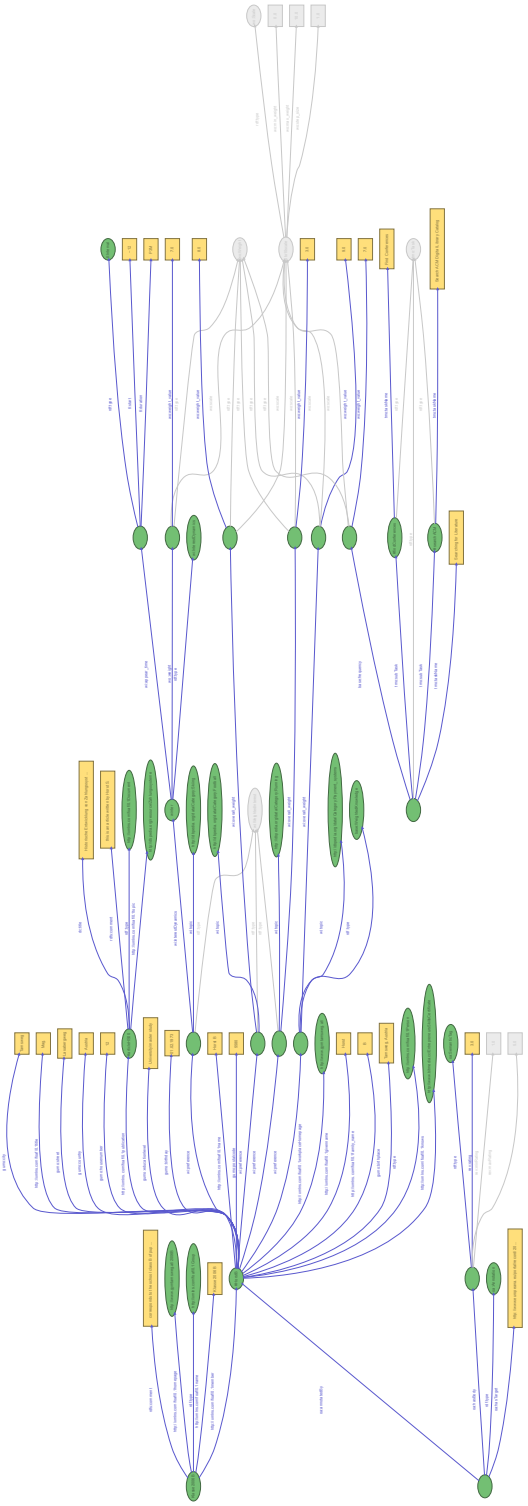

Figure 10: user profile visualization of Horst B.

A.2 Questionnaire for Test Data Acquisition

ProbandInnennummer:

Fragebogen

DEMOGRAPHISCHE ANGABEN: **D**

AA1 ProbandInnennummer:
AA3 Alter:
AA4 Geschlecht:
□ männlich □ weiblich □ keines von beiden
AA5 Beruf:
AA6 Geburtsland:
AA7 Beruf:
AA7 Wenn Sie studieren, was ist ihr Studiengang?
AA8 Erfahrung mit Computern: ExpertIn □ BenutzerIn □ Neuling □
AA9 Wie oft benutzen Sie das Internet: Seltener □ einmal die Woche □ täglich für unter zwei Stunden □ öfter □
AA10 Mit welchem Gerät sind Sie hauptsächlich im Internet: Tablet □ Smartphone □ PC/Mac □

ALLGEMEINE FRAGEN: **A**

A1 **Wie schwierig fanden Sie es, geeignete Ressourcen zu finden?**
(sehr schwierig) □ 1 □ 2 □ 3 □ 4 □ 5 (sehr leicht)

A2 **Wie schwierig fanden Sie das Einfügen einer Annotation in eine Webseite?**
(sehr schwierig) □ 1 □ 2 □ 3 □ 4 □ 5 (sehr leicht)

A3 **Wie hilfreich fanden Sie die gefundenen Ressourcen?**
(überhaupt nicht hilfreich) □ 1 □ 2 □ 3 □ 4 □ 5 (sehr hilfreich)

A3 **Wie beurteilen Sie dem Umfang der gefundenen Resourcem?**
(überhaupt nicht umfangreich) □ 1 □ 2 □ 3 □ 4 □ 5 (sehr umfangreich)

A3 **Wie beurteilen Sie die Qualität der gefundenen Resourcem?**
(sehr schlechte Qualität) □ 1 □ 2 □ 3 □ 4 □ 5 (sehr gute Qualität)

A3 **Wie beurteilen Sie die Qualität der Zusatzinformationen zu den Resourcen (Titel, Erstellungsdatum usw.)?**
(sehr schlechte Qualität) □ 1 □ 2 □ 3 □ 4 □ 5 (sehr gute Qualität)

ProbandInnennummer:

FRAGEN ZUR PRIVATSPHÄRE: **P**

P1 Angenommen, es gäbe eine Software, die Ihnen unter Zuhilfenahme Ihrer persönlichen Daten gute, relevante Zusatzinformationen für Webseiten liefern kann. Je mehr Ihrer persönlichen Daten Sie zur Verfügung stellen, desto besser werden ihre Ergebnisse sein. Geben Sie an, welche Ihrer persönlichen Daten Sie für welche Verbesserung der Ergebnisse zur Verfügung stellen würden.

	für perfekte Resultate	für Verbesserung	nie
Vollständiger Name	☐	☐	☐
Organisation (inkl. Abteilung)	☐	☐	☐
Organisation (ohne Abteilung)	☐	☐	☐
Arbeitsfelder	☐	☐	☐
Aufenthaltsort	☐	☐	☐
Komplette Wohnadresse	☐	☐	☐
Stadtangabe in Wohnadresse	☐	☐	☐
Bundesland in Wohnadresse	☐	☐	☐
Land in Wohnadresse	☐	☐	☐
Persönliche Interessen	☐	☐	☐
Kontakte in sozialen Netzwerken	☐	☐	☐
Inhalte in sozialen Netzwerken („Gefällt mir“, „Teilen“, Kommentare“)	☐	☐	☐
Lesezeichen	☐	☐	☐
Browser-History (besuchte Internetseiten)	☐	☐	☐
Zusammenfassung der Browser-History (uni-passau.de statt http://www.rz.uni-passau.de/dienstleistungen-rz/)	☐	☐	☐
Suchhistory (alle verwendeten Suchbegriffe)	☐	☐	☐
Meistgenutzte Suchbegriffe	☐	☐	☐

P2 Wie sensibel in Bezug auf Privatsphäre sind für Sie die folgenden persönlichen Angaben?

	sehr sensibel	sensibel	neutral	wenig sensibel	überhaupt nicht sensibel
Vollständiger Name	☐	☐	☐	☐	☐
Organisation (inkl. Abteilung)	☐	☐	☐	☐	☐
Organisation (ohne Abteilung)	☐	☐	☐	☐	☐
Arbeitsfelder	☐	☐	☐	☐	☐
Aufenthaltsort	☐	☐	☐	☐	☐
Komplette Wohnadresse	☐	☐	☐	☐	☐
Wohnort	☐	☐	☐	☐	☐
Bundesland	☐	☐	☐	☐	☐
Land	☐	☐	☐	☐	☐
Persönliche Interessen	☐	☐	☐	☐	☐
Kontakte in sozialen Netzwerken	☐	☐	☐	☐	☐
Inhalte in sozialen Netzwerken („Gefällt mir“, „Teilen“, Kommentare“)	☐	☐	☐	☐	☐

ProbandInnennummer:

P3 Angenommen, Sie verändern Ihre Privatsphäreeinstellungen in der Software und geben weniger *hochsensible* persönliche Daten preis. Wenn Sie nach dieser Änderung eine Verschlechterung der Resultate feststellen, wann würden Sie die Einstellungen wieder zurück setzen?

□ Nie (wenn eine Information privat ist, gebe ich sie nicht preis, auch wenn das die Ergebnisse verschlechtert)
□ Wenn die Verschlechterung sehr groß ist, Qualität ist nicht ganz so wichtig
□ Schon bei kleinen Verschlechterungen, Qualität ist mir wichtiger
□ Trifft nicht zu, ich würde mich gar nicht um Privatsphäreeinstellungen kümmern.

P4 Angenommen, Sie verändern Ihre Privatsphäreeinstellungen in der Software und geben weniger *leicht sensible* persönliche Daten preis. Wenn Sie nach dieser Änderung eine Verschlechterung der Resultate feststellen, wann würden Sie die Einstellungen wieder zurück setzen?

□ Nie (wenn eine Information privat ist, gebe ich sie nicht preis, auch wenn das die Ergebnisse verschlechtert)
□ Wenn die Verschlechterung sehr groß ist, Qualität ist nicht ganz so wichtig
□ Schon bei kleinen Verschlechterungen, Qualität ist mir wichtiger
□ Trifft nicht zu, ich würde mich gar nicht um Privatsphäreeinstellungen kümmern.

ProbandInnennummer:

Offene Fragen: **O**

B1 Was hätte man besser machen können?

B2 Womit hatten Sie Probleme?

B3 Was fanden Sie gut?

B4 Was ist Ihnen sonst noch aufgefallen? Was möchten Sie uns noch mitteilen?