Aggregated, Interoperable and Multi-Domain User Profiles for the Social Web *

Fabrizio Orlandi
Digital Enterprise Research Institute
National University of Ireland, Galway
fabrizio.orlandi@deri.org

John Breslin
Digital Enterprise Research Institute
National University of Ireland, Galway
john.breslin@nuigalway.ie

Alexandre Passant
Digital Enterprise Research Institute
National University of Ireland, Galway
alexandre.passant@deri.org

ABSTRACT

User profiling techniques have mostly focused on retrieving and representing a user's knowledge, context and interests in order to provide recommendations, personalise search, and build user-adaptive systems. However, building a user profile on a single social network limits the quality and completeness of the profile, especially when interoperability of the profile is key and its reuse on different sites is necessary for providing other types of personalisation. Indeed recent studies have shown that users on the Social Web often use different social networking sites for diverse, and sometimes non-overlapping, purposes and interests. In this paper, we describe our methodology for the automatic creation and aggregation of interoperable and multi-domain user profiles of interests using semantic technologies. Moreover, we propose a user study on different user profiling techniques for social networking websites in general, and for Twitter and Facebook in particular. In this regard, based on the results of our user evaluation, we investigate (i) the accuracy of different methodologies for profiling, (ii) the effect of time decay functions on ranking user interests, and (iii) the benefits of merging different user models using semantic technologies.

Keywords

Social Web, User modeling, Web Personalization, Semantic Web, Web of Data

1. INTRODUCTION

Users on the Social Web interact with each other, create/share content and express their interests on different social websites with many user accounts and different purposes. On each of these systems personal information, consisting of a portion of the complete profile of the user, is recorded. With respect to "complete user profile" we intend the full set of personal information belonging to a person obtained by aggregating the distributed partial user profiles on each Social Web system. Each partial user profile might contain the user's personal and contact information, her interests, activities and social network of contacts. In this work in particular we focus on user profiles of interests as structured and ranked collections of concepts relevant to the users. These details are often used by applications for personalisation and recommendation purposes. All the distributed user profiles on the Web represent different *facets* of the user therefore their aggregation provides a more comprehensive picture of a person's profile [4]. Aggregation of user profiles brings several advantages: it allows for information reuse across different systems, it solves the well-known "cold start" problem in personalisation/recommendation systems, and provides more complete information to each individual Social Web service. However, the aggregation process is a non-trivial problem which derives from the most popular data integration issues: entity matching and duplicates resolution, conflicts resolution, heterogeneity of the data models of the sources and the consequent need of a common target data model are the most important ones. Using standard semantic technologies to represent the data sources would help in solving these issues and it would provide a unified representation of the target data model. Furthermore a complete semantic representation and management of the provenance of user data addresses the duplicate/conflict resolution issues, since it would allow to track the origins of the data at any point of the integration process [10]. Several approaches for aggregating and representing multi-domain user models have been presented in the state of the art (see Section 2) but in most of the cases they are not aimed at defining a standard, source-independent, architecture that allows for interoperability and integration of profiles of interest on the Web of Data. The use of the best Linked Data principles and the integration with the Web of Data is crucial, as it automatically provides a standard "platform" for the representation of the user data with popular vocabularies and enables for semantic data enrichment using the many open datasets on the LOD (Linked Open Data) cloud [6]. At the same time approaches that aim at integrating user models with the Web of Data [17] are system dependent and do not focus on aggregation of user data from different sources.

In this paper we present an architecture for the automated extraction, aggregation and representation of user profiles of interest. The architecture is platform-independent and

*The work presented in this paper is funded in part by Science Foundation Ireland under grant number SFI/08/CE/I1380 (Líon 2), and in part by an IRCSET scholarship.

I-SEMANTICS 2012. 8th Int. Conf. on Semantic Systems, Sept. 5-7, 2012, Graz, Austria.

can be applied to every Social Web system since it is based on the analysis of text produced by the user, *i.e.* through the analysis of messages or other social networking activities such as comments, places checked in, liked links, etc. In particular for our experiment we implemented a system that computes profiles harvested from Twitter[1] and Facebook[2] user accounts. The resulting user profiles consist of entities and concepts representing interests, activities and contexts of the users. We use DBpedia[3] resources and categories to represent the entities included in a profile and we rank the relevance of the interests according to weights computed by three different algorithms described and evaluated in the next sections. The evaluation of the system and the different methodologies has been conducted with a user study with 21 participants. The study reveals interesting results not only for the comparison of the three methods but also for the evaluation of the effects of exponential time decay functions applied to the computation of the weights. Moreover, the difference between the usage of DBpedia resources and DBpedia categories for the representation of the interests and the benefits of the concept expansion obtained using DBpedia are investigated.

The paper is organised as follows: Section 2 summarises the related work, while Section 3 describes a semantic representation of user profiles of interests. In Section 4 we describe the architecture and implementation of the profiling application and the different methodologies proposed for weighting the sets of interests. The evaluation of the system and a user study is described in Section 5, before analysing the results and concluding the paper.

2. RELATED WORK

During the past few years we have assisted to the growth of Web applications using or collecting data on their users and their behaviour in order to provide adapted and personalised contents and services. This caused the need for exchange, reuse, and integration of their data and user models. A new research challenge then emerged, seeking solutions for user modelling and personalisation across application boundaries. Recent relevant studies of the state of the art for this field are described in [8] and in [19], where authors focus on adaptive systems adopting Semantic Web technologies.

Interesting research that combines user information retrieval/profiling and the Semantic Web has been presented by Szomszor et al. [17]. The authors investigate the idea of merging users' distributed tag clouds to build richer profile ontologies of interests, using the FOAF vocabulary and matching concepts to Wikipedia categories. In [7] Carmagnola et al. describe one of the most advanced user modelling systems adopting semantic technologies. The use of RDF for representing user models, and the reasoning capabilities implemented on top of the user models in order to obtain automatic mapping between heterogeneous concepts, are the strongest points of their implementation. Moreover, an extensive approach for ontology-based representation of user models was proposed by Heckmann et al. by introducing GUMO [11], a General User Modeling Ontology for the uniform interpretation of distributed user models.

As regards user profiling on social networks, the work presented in [18] and [1] shows an interesting and similar approach for creating RDF-based user profiles on Twitter according to the frequency of the entities extracted from the user's tweets. The profiles are then modelled primarily using the FOAF vocabulary. Particularly relevant is the fact that the authors demonstrate the benefits of the amalgamation of multiple Web 2.0 user-tagging histories in building personal semantically-enriched profiles of interest. However, as a comparison, in our work we focus more on investigating different profiling methods (especially using DBpedia categories) and evaluate the effects of time decay functions through a user study. An analysis of different temporal patterns and dynamics for Twitter user profiles is also provided by the same authors in [2].

Recent related work has also been published in [16] where the authors describe a system for people recommendation based on *User Interaction Profiles* built extracting entities and keywords from user posts on social networks (from Twitter, in their experiment). A similar architecture for the generation of the profiles is proposed and disambiguation and concept expansion is also done using DBpedia and semantic technologies. On the other hand, an evaluation of the system and the profiling algorithm is not provided and temporal features of user posts are not considered. Despite the interesting combination of traditional content analysis techniques with semantic technologies, in this work the focus is more on building a framework for people recommendation during web navigation.

Relevant related work on Semantic Web applied to user modelling and personalisation has been done by Aroyo et al. [5]. In this work the authors highlight the challenges they see in the near future for user modelling and the adaptive Semantic Web and a review of the research in this field is provided. In the state of the art review the authors analyse the differences between past user modelling solutions (in traditional "closed world" Web-based or application-based systems) and new research on "open" and Semantic Web based solutions. The fundamental tasks identified by the authors that contribute to user modelling are: user identification, user property representation, and sharing adaptation functionalities. The major question in user identification investigates how to identify a person on the Web, her multiple identities across different applications and what are the trust and privacy aspects involved. As regards user knowledge the main challenge is to find ways to share user models, and this implies the definition of common vocabularies and interoperable representations of objects and values of user properties. Finally in their work Aroyo et al. highlight an important aspect about the openness of the Web of Data and the related implications of this on users' experience: an open approach to user knowledge would produce different new use cases and knowledge management approaches, especially users should then be able to inspect and edit their own data (*scrutability* of user profiles). Related and more practical work by the same authors and others is described in [15] where, as part of the NoTube project, by using the Linked Data cloud, semantics can be exploited to find complex relations between the user's interests and background information of TV programmes, resulting in potentially interesting recommendations. Also in another paper [9] Denaux et al. present how interactive user modelling and adaptive content management on the Semantic Web can be integrated

[1] `https://twitter.com`
[2] `https://www.facebook.com`
[3] `http://dbpedia.org/`

in a learning domain to deal with common adaptation problems (e.g. cold start, inaccuracy of assumptions, knowledge dynamics, etc.).

To note also that some of the systems for user model interoperability implement their reasoning capabilities on top of the user data not using Semantic Web technologies but using non-standard application-specific algorithms, making interoperability with other systems more difficult to achieve.

3. INTEROPERABLE AND MULTI-DOMAIN USER PROFILES

The steps involved in our framework for the extraction and generation of user profiles from social networking websites can be summarised with the following main stages. *First*, the data extraction from each specific social networking service and the subsequent generation of application-dependent user profiles. After this phase the *next* steps involve the representation of the user models using popular ontologies, and then, *finally*, the aggregation of the distributed profiles. In this section we describe our RDF modelling solution for multi-domain user profiles of interest and we detail how we integrate user data with the Web of Data and in particular DBpedia. Semantic Web technologies and standard ontologies are the main supports for the development of interoperable services, and these standards make it easier to connect distributed user profiles.

3.1 Representing User Profiles of Interest

A possible and popular modelling solution for profile data is the generation of profiles described using the FOAF vocabulary[4]. FOAF is one of the most popular lightweight ontologies on the Semantic Web and using this vocabulary as a basis for representing users' personal information and social relations eases the integration of heterogeneous distributed user profiles. An important part of a user profile is represented by the user interests. In this work we focus in particular on this part of a profile, on how to automatically retrieve interests from social networking sites and how to compute weights expressing their relevance. In Listing 1 we display an example of an interest about "Semantic Web" with a weight of 0.5 on a specific scale (from 0 to 1) using the Weighted Interests Vocabulary (WI)[5] and the Weighting Ontology (WO)[6]. In order to compute the weights for the interests common approaches are based on the number of occurrences of the entities, their frequency, and possibly some additional factors. These factors might depend on whether or not the interest was implicitly mined or explicitly showed by the user, or depending on a time-based function which computes the decay of the interests over time, or based on the trustworthiness of the social platform, etc. In Section 4 we describe the different weighting schemes and algorithms adopted and evaluated in this work.

3.2 Leveraging Provenance of User Data

Provenance of data is important in this context as it allows data consumers to understand the origins of the interests (time- and source-wise) which are the result of an integration process. Some data consumers might want to give more relevance to some data sources rather than others according to particular trust measures or differences of contexts and use cases. Moreover it would be possible to recompute new aggregated weight values based on different weighting-schemes and the original data, or enforce privacy rules on the user data based on particular preferences. As regards provenance of the interests, as showed in Listing 1, we use the property `wasDerivedFrom` from the Open Provenance Model[7] (OPM) to state that the interest was originated by a specific user account on a website. In the example in Listing 1 we can observe that the interest is derived from both Twitter and Facebook user accounts.

```
<foaf:topic_interest rdf:resource="http://dbpedia.
    org/resource/Semantic_Web" />
<wi:preference>
  <wi:WeightedInterest>
    <wi:topic rdf:resource="http://dbpedia.org/
        resource/Semantic_Web" />
    <rdfs:label>Semantic Web</rdfs:label>
    <wo:weight>
      <wo:Weight>
        <wo:weight_value rdf:datatype="http://www.
            w3.org/2001/XMLSchema#double">0.5</wo:
            weight_value>
        <wo:scale rdf:resource="http://example.org
            /01Scale" />
      </wo:Weight>
    </wo:weight>
    <opm:wasDerivedFrom>
      <sioc:UserAccount rdf:about="http://twitter.
          com/BadmotorF">
      </sioc:UserAccount>
    </opm:wasDerivedFrom>
    <opm:wasDerivedFrom>
      <sioc:UserAccount rdf:about="http://www.
          facebook.com/fabriziorlandi">
      </sioc:UserAccount>
    </opm:wasDerivedFrom>
  </wi:WeightedInterest>
</wi:preference>
[...]
<wo:Scale rdf:about="http://example.org/01Scale">
  <wo:max_weight rdf:datatype="http://www.w3.org
      /2001/XMLSchema#decimal">1.0</wo:max_weight>
  <wo:min_weight rdf:datatype="http://www.w3.org
      /2001/XMLSchema#decimal">0.0</wo:min_weight>
</wo:Scale>
```

Listing 1: Representing an interest (Semantic Web) and its weight (0.5) extracted from two sources

3.3 Interests on the Web of Data

An important aspect is the use of DBpedia to represent the interests of the users. DBpedia[8] is the semantic representation of Wikipedia. Thanks to its large dataset (around 1 billion RDF triples) and its cross-domain nature DBpedia has become one of the most important and interlinked datasets on the Web of Data. Representing interests using DBpedia resources has two main advantages: integrates the user profiles with the Linked Data cloud, and provides a larger and "fresher" set of terms as compared to traditional taxonomies or lexical databases such as WordNet[9]. In [13] the authors analyse the benefits of using Wikipedia (or DBpedia) for computing semantic relatedness and for named entity representation as compared to WordNet and other knowledge bases. In this work we use DBpedia not only to link to its entities but also to extract related categories for concept expansion and to analyse the structure of

[4] `http://www.foaf-project.org/`
[5] WI Specification: `http://purl.org/ontology/wi/core#`
[6] WO Specification: `http://purl.org/ontology/wo/core#`
[7] OPM Specification: `http://openprovenance.org/`
[8] `http://dbpedia.org/`
[9] "About WordNet" 2010. `http://wordnet.princeton.edu`

the categories graph in order to understand the relevance of a category for representing a user interest. Our plan is to extend this analysis also to other Linked Data sources.

4. AUTOMATED AGGREGATION OF USER PROFILES

This section provides a description of the architecture proposed for the automated creation and aggregation of interoperable and multi-source user profiles (Sec. 4.1). We also detail the experiment we conducted in order to evaluate our architecture and the different approaches for ranking user interests (Sec. 4.2). A complete analysis of the experiments and a user study is provided in Section 5. The experiments have been conducted using Facebook and Twitter as sources.

4.1 Software Architecture

We implemented a web service (written in PHP) that requires users to log-in with two of their Social Web user accounts and returns a representation of their user profile of interests in RDF. The generated profile is the aggregated result of the analysis of their activity on such services. From an architectural perspective, the profiling framework is composed of three main modules (Fig. 1):

(1) Service-specific data collector;

(2) Data analyser and profile generator;

(3) Profiles aggregator.

The second and third modules include the representation of the profiles of interests using the modelling solution described in the previous Section 3. In *module (2)* the semantic representation involves only a specific single source profile, while in *module (3)* RDF is generated for the final aggregated user profile. The implementation of the system for our experiment and evaluation is based on two of the most popular social networking sites: Facebook and Twitter.

4.1.1 Service-specific data collector

The first module is the module that interacts directly with the source of the profile, the social networking site. This module is responsible for the interaction with the service API, the user authentication, and the data collection from the API. In order to collect private data about users on social websites it is necessary to have access granted to the data by the users. Then it is necessary to request access to the profile data in order to fetch most of the data which is often private by default. In particular, dealing with Facebook and Twitter, we implemented the OAuth 2.0 [10] authentication system required by these platforms to access users' private data. We implemented two distinct modules, one for each social service, each of them including the OAuth authentication system. We adopted two different libraries for PHP: *Twitter-async*[11] and *Facebook PHP-SDK*[12]. Using the Twitter API we are able to request up to 3,200 of a user's most recent statuses, while Facebook adopts rate limits. The type of data we collected from Facebook is: status messages posted on the user's wall, the entities liked, the places checked-in and user profile information. In the same way on Twitter we retrieve the status messages posted by the user on his/her timeline and other users' messages that the user "retweeted".

4.1.2 Data analyser and profile generator

Once the user data has been collected from the different platforms the next step is the analysis of the data in order to identify entities and generate the profiles. In this work we use a named entity recognition software to extract entities from the text retrieved at the previous stage. In particular we use Zemanta[13], a web service that exposes an API and provides text analysis tools to developers. The service in particular offers natural language processing capabilities and a named entity extractor that spots entities such as places, persons, organisations, etc. and provides the related DBpedia resources. It performs entity disambiguation, as entities are linked to URIs on the Linked Data cloud and ambiguities are resolved analysing the context of the sentences[14]. We chose Zemanta for its automated DBpedia URIs suggestion capabilities and for its satisfying performances in analysing short messages such as tweets. According to the state of the art Zemanta, in comparison with similar services such as Alchemy API[15], DBpedia Spotlight[16] and Open Calais[17], performs slightly better than the others. Recent research on this topic [14] is supporting this statement and suggests Alchemy API and DBpedia Spotlight as the main alternatives. According to the study Zemanta has higher precision than the other tools in recognising named entities and disambiguating them with proper URIs (which is the most important feature for our work). This is supported by a substantial agreement between the evaluators during the experiments conducted by Rizzo et al. According also to other studies [12] Zemanta dominates in precision but has lower recall than DBpedia Spotlight and the WikiMachine[18] that have similar F_1-scores. To note also that other tools such as Alchemy API perform better in categorisation but this feature is not required in our work since we can use the DBpedia taxonomy for this task. For an extensive evaluation of these tools we rely on the work published in [14] and [12] as this is not the focus of our work.

In our framework in particular we perform entity extraction algorithm on every message and social activity that we collected at the previous stage. For each message we then record the time the action was performed by the user and the set of entities retrieved for that message. A list of entities (DBpedia URIs provided by Zemanta) is then populated during this phase. For every entity we record the number of occurrences and the timestamps for each of them. Hence, not only the latest occurrence is kept into memory, but also the timestamps for all the previous ones. This part is important for computing the weights of the interests.

In this regard we combine the number of occurrences with a time decay function that evaluates the distribution over time of the interests. We use an exponential decay function to evaluate the relevance of each interest according to its position on the user timeline. The function gives higher

[10] http://oauth.net/2/
[11] https://github.com/jmathai/twitter-async
[12] https://github.com/facebook/php-sdk
[13] http://developer.zemanta.com/
[14] Zemanta API companion documentation: http://developer.zemanta.com/docs/
[15] http://www.alchemyapi.com/
[16] http://dbpedia.org/spotlight
[17] http://www.opencalais.com/
[18] http://thewikimachine.fbk.eu

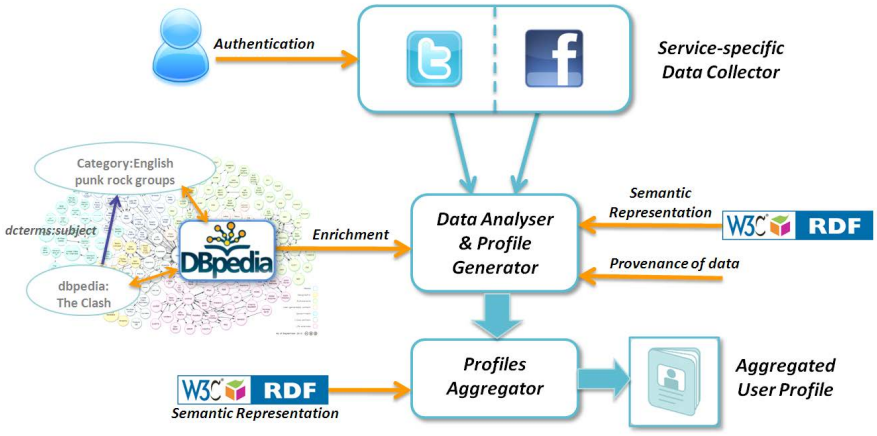

Figure 1: Architecture Diagram

weight for interests occurred recently and lower for older interests. The exponential decay function is:

$$x(t) = x_0 e^{-t/\tau} \quad (1)$$

Where: $x(t)$ is the quantity at time t, $x_0 = x(0)$ is the initial quantity (at time $t = 0$), $\tau = 1/\lambda$ is a constant called *mean lifetime* and λ is a positive number called the *decay constant*.

Applying this function to our use case, in order to compute the time decay of the interests, we need to arbitrarily choose values for x_0 and τ which are constants of the function. For our experiment we set $x_0 = 1$, the maximum possible value of the function. We also defined an initial time window where the interests are not discounted by the decay function (7 days). Moreover, in order to identify an appropriate value for τ, we decided to choose two possible values and evaluate them with an experiment and a user study. The constant τ represents the time at which the function value is reduced to $1/e = 0.368$ times its initial value x_0. In our experiment we evaluate the following two values: $\tau = 120 days$ and $\tau = 360 days$. From a practical point of view the two values indicate that an interest value is discounted to 37% of its initial value respectively after 120 and 360 days.

The exponential decay function is directly applied to the frequency value of the interests, calculated as the ratio between the number of the interest occurrences and the total number of occurrences of all the interests. As regards the time considered for the decay function (the value of t) we compute the average time of the timestamps for each interest. Following the computation of the weights for all the interests, all the values are then normalised in an interval between 0 and 1. Finally, the set of interests generated after this second phase has to be represented in RDF according to the modelling solution described in Section 3.

4.1.3 *Profiles aggregator*

The final phase of the profiling framework is the aggregation of all the single source user profiles. The problem that arises when merging user profiles is the necessity to resolve shared interests reoccurring on different profiles and to recalculate a global weight for these interests. Their new aggregated weight should then be higher than their weight on a single profile, as reoccurring concepts on different social media sites indicate a strong interest. If the same interest is present on two or more profiles it is necessary to: represent the interest only once, compute its new global weight, and update the provenance of the interest keeping track of the sources where the interest was derived from. As regards the computation of the aggregated global weight for the interest generated by multiple sources, we propose a simple generic formula that can be adopted for merging the interest values of many different sources. The formula is as follows:

$$G_i = \sum_s W_s * w_{is} \quad (2)$$

Where: G_i = global weight for interest i, W_s = weight associated to the source s, w_{is} = weight for the interest i in source s.

Using this formula it is possible to specify static weights associated to each source depending on which source we want to give more relevance to. In our particular experiment we did not assign different weights to Twitter and Facebook. We considered every social website equally in terms of relevance, hence we multiply each of the two weights by a constant of 1/2 and then we sum the results. The following formula summarises the computation of a new global weight (G) as result of the two original weights (W_1, W_2). It is the same formula that we propose in the previous section (formula 2) with the following values: $W_s = 1/2 \; \forall s$. Hence: $G_i = 1/2 * w_{i1} + 1/2 * w_{i2}$.

The different values associated to W_s depend on the particular source but can also be associated to a type of source. For example microblogging platforms (e.g. Twitter, Identica, etc.) could be associated with the same value. This fine-grained weighting strategy is dependent on the particular application for the user profiles or on the users themselves and we plan to investigate it further in our future work.

4.2 Experiment

This section describes the experiment that has been conducted in order to evaluate the implementation of the system and different aspects and methodologies of user profiling. The first aim of this experiment is to evaluate the accuracy of aggregated user profiles in relation to the weighting-scheme and the ranking of the interests. The system allows users to generate user profiles from their Twitter and Face-

book user accounts. In particular we generate 6 types of user profiles which differ for the following aspects: (i) The type of DBpedia entities adopted (either Categories or Resources). (ii) The type of weighting-scheme for category-based methods (two different methods). (iii) The type of exponential decay function (either with a shorter time decay parameter $\tau = 120$ days, or a longer one $\tau = 360$ days). As regards the first aspect, the category-based methods implemented extract from DBpedia all the related categories of the DBpedia resources that have been computed with the resource-based methods. As soon as we get a DBpedia entity from the entity recognition tool, this takes part of the resource-based profile. Then, for every resource collected we query the DBpedia SPARQL endpoint[19] to retrieve the categories that are connected to the resources. A DBpedia resource is linked to its categories through the Dublin Core[20] `subject` property. From each category, which is defined as a `skos:Concept`[21], is also possible to navigate the categories graph to obtain more related categories using the `skos:broader` and `skos:narrower` relationships. This option, not implemented with this experiment, would be useful for use cases where is necessary to broaden the user profiles, for instance for recommendation systems. Once all the categories are retrieved from DBpedia starting from the original resource-based user profiles, we can create the category-based profiles and assign different weights to the categories according to different weighting-schemes. This involves then the second aspect targeted by our experiment.

We developed two different weighting-schemes for the categories. The first one is the most straightforward one: it propagates the weights of the resources computed for the resource-based method to the categories. Hence, the weight of each category is the sum of all the weights of the interests/resources belonging to that category. The idea of the second type of weighting-scheme is to reduce the weight of the category (computed in the same way as the first weighting-scheme) if the category is a too "broad" or generic category, so it is not descriptive for a user profile. More in detail, analysing the structure of the categories on DBpedia we noted that generic categories usually contain many resources or have several subcategories. We then implemented a solution to lower the weight of this type of category. In this case the discount value that multiplies the original weight of the category is computed as follows:

$$CategoryDiscount = \frac{1}{log(|SP|)} \cdot \frac{1}{log(|SC|)} \qquad (3)$$

where: SP = *Set of Pages belonging to the Category*, SC = *Set of Sub-Categories*.
The number of subcategories and pages is retrieved again using the DBpedia SPARQL endpoint. This method, for example, discounts the value of too generic categories such as *"Living People"*, which are not meaningful and representative of a user interest. At the same time the method keeps the original weight for relevant and particular categories such as *"RDF"*.

The third and last aspect we chose for our experiment is the exponential time decay function applied to the computation of the weights. As explained in Section 4.1 we chose two values of time decay parameter (120 and 360 days) and implemented all the three different methods two times with the two decays. Hence, in conclusion, for each user we ran our experiment with 6 different profiling algorithms: *resource*-based profiling, *category*-based profiling *1st* method and *category*-based profiling *2nd* method, each of them twice because of the two time decay parameters (we use the following abbreviations: *Res 360, Res 120, Cat1 360, Cat1 120, Cat2 360, Cat2 120*). The generation of the 6 user profiles takes from 6 to 9 minutes on a standard dual core laptop. In Table 1 we display the average number of interests generated for each method. This has been evaluated with 21 users (see Sec. 5) and, on average, using category-based methods generates 6.8 times more interests than the resource-based ones, and the longer time decay (3 times longer) generates 1.4 times more interests.

Res 360	**Res 120**	**Cat 360**	**Cat 120**
44.5	33.1	308.1	221.8

Table 1: Average # of interests per profiling method

5. ANALYSIS AND EVALUATION

-	**Facebook**	**Twitter**
every day	**66.7%** (14 users)	14.3% (3 users)
every other day	19.0% (4 users)	14.3% (3 users)
once/twice a week	9.5% (2 users)	23.8% (5 users)
once every two weeks	0.0% (0 users)	**28.6%** (6 users)
once a month	4.8% (1 user)	19.0% (4 users)

Table 2: Active usage of Facebook and Twitter

In this section we analyse the evaluation of the implemented system and the different methodologies proposed. In order to evaluate the validity of our approach for generating aggregated user profiles we conducted a user study with 21 users. Demographics include users from 21 to 45 years old, all of them proficient with Social Web systems and 76% of them working/studying in information technology fields. The survey we proposed to the users is composed of 10 questions and the average time taken by the users to complete it was between 9 and 10 minutes. Table 2 shows their answers for: "How often do you *actively* use Facebook/Twitter? (i.e. post a message/link, press "like" buttons, check-in, etc.)". From the table is clear that in general our sample uses more actively Facebook than Twitter.

The second type of question we asked users was about enumerating a list of entities and concepts that they were expecting to be representative of their interests, activities and context on both Twitter and Facebook. This question helps understanding if the topics expected by the users are represented also in the user profiles that we generated. Using the answers to these questions we were able to identify the interests in the generated profiles that were relevant to the users and the interests that were expected by the users but missing in our profiles. This allowed us to compute an approximate *recall* value for our profiles, even though this

[19] `http://dbpedia.org/sparql`
[20] `http://dublincore.org/documents/dcmi-terms/`
[21] `http://www.w3.org/2004/02/skos/core.html`

-	Cat1 360	Cat1 120	Cat2 360	Cat2 120	Res 360	Res 120	Baseline
Average Score	5.67	5.20	5.49	5.26	7.24	6.81	3.46
# of Non-Relevant	31	42	34	46	21	22	74
Tot # of Scores	210	209	209	210	210	205	200
Precision	0.857	0.799	0.837	0.781	0.900	0.893	0.630
MRR	0.921	0.937	1.00	0.933	1.00	1.00	0.858
P@10	0.852	0.800	0.838	0.781	0.900	0.895	0.610

Table 3: Statistics about the user study for each of the 6 profiling methods and the baseline.

method might not be very accurate since users had no restrictions in choosing their expected interests. Also, since users do not have perfect memory, we acknowledge the fact that this recall measure is just an estimation and an accurate recall value in this case cannot be computed. The computed average recall value for all the profile types is: 0.740. Next we evaluate the precision according to different measures.

The other remaining 6 questions were all similar, and required users to give a relevance score to each of the top 10 interests for each of the 6 proposed profiling methods. For each method we provided a table of ten interests ordered by weight. These methods have been previously described in Section 4.2. The exact question formulated to the users was: "Consider Table X. Please rate how relevant is each concept for representing your personal interests and context.". The options available to users for rating the interests were the following: *0 (not at all or don't know), 1 (low relevance), 2, 3, 4, 5 (high relevance)*. Users were then rating the interests on a scale from 0 to 5 (rescaled then in values between 0 and 10) and they were supposed to give a score equal to 0 in case the interest was totally unrelated or unknown.

In Table 3 we summarise the values obtained for each of the 6 methods considering as a non-relevant result the case when the rate value is 0 (so the non-relevant value is below 2 in a 0 to 10 scale). The values of the average score are on a 0 - 10 scale. We use the Mean Reciprocal Rank (MRR) and the Precision at $K = 10$ (P@10) statistical measures to evaluate the accuracy of the profiles and the ranking/weighting scheme. In Figure 2 we can see a comparison of the MRR and P@10 values calculated both for the case that considers non-relevant an interest with score lower than 2, and for the other which considers non-relevant scores lower than 4.

As we can see all the values of MRR and P@10 are satisfying and encouraging. As a comparison with traditional non-semantic approaches we also included an evaluation of a "*Baseline*" method (Table 3). This method is a simple traditional approach that retrieves the most frequent words from the user posts and ranks them according to their number of occurrences. Stemming is applied in our case and stopwords are also removed. As showed in Table 3 this method performs clearly worse than all the other "semantically enhanced" methods and the evaluation has been completed by almost all the users who evaluated also the other methods.

Further, the two methods using DBpedia resources, and not the categories, perform better than the others using categories, and at the same time the results for $\tau = 360days$ are slightly better than for $\tau = 120days$. Therefore we would infer that a longer time frame, and a smoother exponential decay function, would better represent users' interests. To note that this is probably true in cases similar to this one, where the aim of the profile is to globally represent user interests and contexts, but it might not be true in cases such as news recommendations where a "fresher" and updated user profile might perform better (see the work in [3]). As regards the statistical significance of the results, we tested our data with both a Wilcoxon's matched pairs test and a paired two-tailed t-Test. The first one is more appropriate because it is a non-parametric method and also our sample is relatively small. Yet, both the tests provide the same results. We tested the differences between the three main methods (and especially the differences between category-based methods and resource-based ones) we calculated p values lower than 0.05, which confirms the significance of the results. As regards the comparison between the samples with two different τ values the Wilcoxon's test rejected the hypothesis of statistical significant difference between the two samples but the computed p values are very close to the α value ($\alpha = 0.05$). This means that we cannot state that the results for those cases are significant, although those numbers are not very high. However, we probably have to increase the number of users in the sample in order to test whether the theory is statistically valid or not.

Interesting to note that DBpedia resources are slightly more precise and specific for building profiles than the related categories extracted, however the results obtained using categories are very close to the traditional methods using just DBpedia resources. Moreover, as an advantage for using categories, as we have shown in Section 4.2, the number of categories that can be extracted for profiling a user is almost 7 times larger than the number of resources. This is particularly useful for recommendation use cases, where there is a need of getting as much related concept as possible for profiling a user. Further, according to the results, we think that mixed approaches adopting both categories and resources for user profiling can be highly beneficial and need to be investigated. According to users' feedback during the survey, DBpedia resources revealed to be often very specific and narrow, so not always appropriate for representing user interests. On the contrary, the categories for the first method were sometimes too generic (e.g. the frequently occurring *"Category:Living_People"*) and although the second category-based method is capable of removing the very broad categories from the top of the interests' list, it has the problem of introducing more noise.

6. CONCLUSION AND FUTURE WORK

In this paper we described the architecture of a user profiling framework and a methodology for the aggregation of different user profiles on the Social Web. We implemented the architecture with a system that generates user profiles of interest from both Twitter and Facebook and we evaluated the validity of different approaches for user profiling through

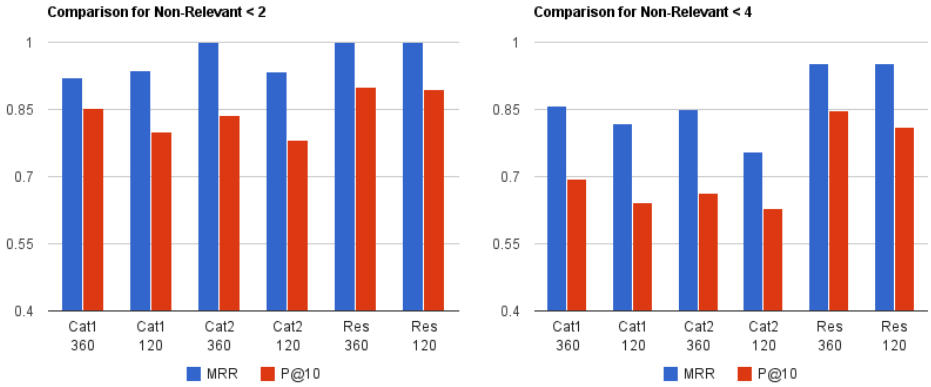

Figure 2: User Evaluation - MRR and P@10

a user study. From the evaluation emerged that, despite the overall good quality of the results, user profiles based on interests described using DBpedia resources are slightly more accurate than those built using DBpedia categories. However the difference between the two approaches is narrow and the second approach has the advantage of producing almost 7 times more user interests. This can be highly beneficial for recommendation systems for example. The potentialities of an exponential time decay function applied to the computation of the interests' weights have also been studied. As a result we found that, for the generic representation of user interests, a slower decay function produces more accurate profiles, as it does not penalise excessively interests occurring also in the past and provides a more complete picture of the user. Then, a second method for the computation of category-based profiles has been proposed and evaluated but, according to the user study, it requires some more improvements. Indeed, it helps in filtering out categories that are too generic but it also introduces some undesired noise in the profiles. In the near future we plan to improve the second method for category-based profiles, then we would like to consider mixed approaches combining the resource-based user profiles with the category-based ones. We also plan to apply clustering algorithms to the profiles in order to automatically group similar interests.

7. REFERENCES

[1] F. Abel, Q. Gao, G. Houben, and K. Tao. Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. In *ESWC 2011*.

[2] F. Abel, Q. Gao, G. Houben, and K. Tao. Analyzing Temporal Dynamics in Twitter Profiles for Personalized Recommendations in the Social Web. In *ACM WebSci'11*, 2011.

[3] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing User Modeling on Twitter for Personalized News Recommendations. In *UMAP 2011*.

[4] F. Abel, N. Henze, E. Herder, and D. Krause. Interweaving Public User Profiles on the Web. In *User Modeling, Adaptation, and Personalization*, 2010.

[5] L. Aroyo and G. Houben. User modeling and adaptive Semantic Web. *Semantic Web Journal*, 2010.

[6] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3), 2009.

[7] F. Carmagnola. Handling Semantic Heterogeneity in Interoperable Distributed User Models. *Advances in Ubiquitous User Modelling*, 2009.

[8] F. Carmagnola, F. Cena, and C. Gena. User model interoperability: a survey. *User Modeling and User-Adapted Interaction*, 2011.

[9] R. Denaux, V. Dimitrova, and L. Aroyo. Integrating open user modeling and learning content management for the semantic web. *User Modeling*, 2005.

[10] O. Hartig and J. Zhao. Publishing and Consuming Provenance Metadata on the Web of Linked Data. In *3rd Int. Provenance and Annotation Workshop*, 2010.

[11] D. Heckmann, T. Schwartz, B. Brandherm, M. Schmitz, and M. von Wilamowitz-Moellendorff. Gumo, the general user model ontology. In *User Modeling 2005*, 2005.

[12] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia Spotlight: Shedding Light on the Web of Documents. In *I-Semantics 2011*, 2011.

[13] S. P. Ponzetto and M. Strube. Knowledge derived from Wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, 2007.

[14] G. Rizzo and R. Troncy. NERD: Evaluating Named Entity Recognition Tools in the Web of Data. In *ISWC'11 - Workshop on Web Scale Knowledge Extraction*, 2011.

[15] B. Schopman and Others. NoTube: making the Web part of personalised TV. In *ACM WebSci'10*, 2010.

[16] J. Stan, P. Maret, and V. Do. Semantic User Interaction Profiles for Better People Recommendation. *Conference on Advances in Social Networks Analysis and Mining*, 2011.

[17] M. Szomszor, H. Alani, I. Cantador, K. O'Hara, and N. Shadbolt. Semantic modelling of user interests based on cross-folksonomy analysis. *ISWC*, 2008.

[18] K. Tao, F. Abel, Q. Gao, and G. Houben. TUMS: Twitter-based User Modeling Service. In *International Workshop on User Profile Data on the Social Semantic Web (UWeb)*, 2011.

[19] I. Torre. Adaptive systems in the era of the semantic and social web, a survey. *User Modeling and User-Adapted Interaction*, 2009.