Users as crawlers: exploiting metadata embedded in Web pages for user profiling

Dario De Nart, Carlo Tasso, Dante Degl'Innocenti

Artificial Intelligence Lab
Department of Mathematics and Computer Science
University of Udine, Italy
{dario.denart,carlo.tasso}@uniud.it, dante.deglinnocenti@spes.uniud.it

Abstract. In the last years we have witnessed the rapid growth of a broad range of Semantic Web technologies that have been successfully employed to enhance information retrieval, data mining and user experience in real-world applications. Several authors have proposed approaches towards ontological user modelling in order to address different issues of personalized systems, such as the cold start problem. In all of these works, non-structured data such as tags are matched, by means of various techniques, against an ontology in order to identify concepts and connections between them. However, due to recent popularity of semantic metadata formats such as microformats and RDFa, structured data are often embedded in many Web contents, with no need to "guess" them using a support ontology which may not be coherent with the actual content and the original goals of the author. In this paper we propose a novel approach towards ephemeral Web personalization based on extraction and enrichment of semantic metadata embedded in Web pages. The proposed system builds, at client-side, a rdf network that can be queried by a content provider in order to address personalized content.

Key Words: User Modeling, Semantic Web, Ephemeral Personalization, RDFa

1 Introduction

Personalization is one of the leading trends in Web technology today and we have all stumbled upon it in a way or in another while surfing the Web. Most of the times the process is evident, for instance when web sites require us to sign in and ask for our preferences in order to maintain an accessible user profile. But in other cases personalization is more subtle and is hidden to the user

Ephemeral personalization [9], for instance, aims at providing personalized content fitting only short-term interests that can expire after the navigation session. Most of the times there is no need for the user to sign in order to exploit ephemeral personalization, since all the information needed to determine which content should be presented may be found in his/her browsing cache and/or content providers are not interested in modelling and archive such short-term interests. An example of ephemeral personalization is targeted adversing, that is providing personalized ads to users as they browse the Web. This task is currently accomplished by checking which cookies are present in the client's browser cache and selecting candidate ads accordingly. This process, however, in most cases results in a particular ad from a previously visited site, "stalking" in such way the user throughout all his/her browsing activities. As the authors of [7] suggest, this may generate a revenue for the advertiser by encouraging customers to return, but can also be extremely annoying and the users may perceive their privacy attacked. Other forms of ephemeral personalization are guided by contextual information derived from the IP address of the client or by analysing the content of the pages that the client requests, like in Amazon's product pages, however these are very shallow forms of personalization and do not involve an explicit and persistent user model.

In this work, we claim that there is another way to address ephemeral personalization that, to the best of our knowledge, has never been explored yet. Our approach consists in collecting semantic metadata contained in visited web pages in order to build a client-side user model to be queried by content providers. By doing this the user has total control over his/her user model and the content provider does not need to save and maintain user profiles, therefore privacy risks are significantly reduced.

Before proceeding forth into the technical matter we would like to point out that our approach heavily relies on the availability of semantic metadata embedded in Web pages: the more metadata available, the more detailed the user profile will be; vice versa, if visited pages do not contain metadata, no user profile can be built. Luckily, according to a recent study [2], a huge number of Web sites actually provides semantic annotations, consisting of Microformats, Microdata, or RDFa data, mostly conformed to Facebook's Open Graph metadata protocol[1], hCard, or the Schema.org[2] vocabulary. Since its announcement in 2010, Open Graph caused many concerns about its privacy and service-dependency issues [14], but however, it is greatly contributing to link the Web together. After its adoption by several major players of the WWW, including Google and all its related services, it has affirmed as a de facto standard for RDFa metadata, it has been integrated in CMSs such as Drupal, and can be found in almost any noteworthy site.

The rest of the paper is organized as follows: in Section 2 we briefly introduce some related works; in Section 3 we present our system; in Section 4 we illustrate our data model; in Section 5 we discuss some experimental results and, finally, in Section 6 we conclude the paper.

[1] http://ogp.me/

[2] http://schema.org/

2 Related Work

Several authors have already addressed the problem of generating, parsing, and interpreting structured metadata embedded in Web sites. Automatic metadata generation has been widely explored and can be achieved in many ways: extracting entities from text, inferring hierarchies from folksonomies [13], or exploiting external structured data [8]. Interoperability issues among various metadata formats have been discussed as well: for instance, the authors of [1] propose a metadata conversion tool from microformats to RDF.

Other authors have discussed how Semantic Web tools, such as ontologies and RDF, can be used to model users' behaviours and preferences in Recommender Systems [5]. However, the field on which most research efforts are focused is Personalized Information Retrieval. For instance in [12] is presented an approach towards Ontological Profile building exploiting a domain ontology: as the user interacts with the search engine, interest scores are assigned to concepts included in the ontology with a spreading activation algorithm. The authors of [4] discuss a system that builds a user model aggregating user queries raised within a session and matching them with a domain ontology. Finally, the authors of [3] and [10] suggest that ontological user models can be built as "personal ontology views", that are projections of a reference domain ontology deduced by observing user interest propagation along an ontology network. However, in all these works, user profiles are specializations or projections of a domain ontology and therefore their effectiveness relies on the availability, scope, and quality of such asset.

A recent patent application [15] also claims that the so-called targeting advertising can greatly benefit from the use of semantic user models extracted from Web usage data. The authors, however, do not provide any hint on their extraction technique, focusing, instead, on the architecture and deployment issues of their system. Though many authors have discussed the issues above mentioned, no one, to the best of our knowledge, has ever discussed how to exploit semantic metadata for building personalized interest profiles.

3 System Architecture

In order to support our claims, we developed an experimental system consisting in a client and a server module built using well-known open source tools such as Apache Jena and Semargl. Figure 1 shows the workflow of the system. The basic idea behind our work is that user interests can be identified by observing browsing activity and by analysing the content of visited Web sites, thus our goal is to exploit the user himself as an intelligent Web crawler to provide meaningful data for building his/her personal profile, therefore the project was named *Users As*

Crawlers (herein *UAC*). A compact OWL2 ontology, herein referred as *UAC ontology*, was developed as well in order to introduce new modelling primitives and to allow classification of instances. Among others, the primitives defined in the UAC ontology are: relatedTo, which associates Web pages with DBpedia entities named in the metadata, nextInStream, which associates a page with the next one visited by the user, and previousInStream, which is the inverse of nextInStream.

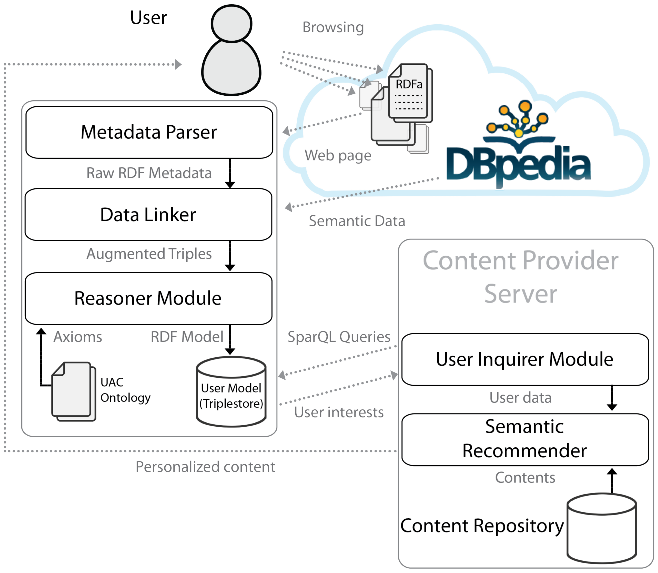


Figure 1. Work flow of the System.

The client module is made of a *Metadata Parser*, a *Data Linker* module, a *Reasoner Module*, and a compact triplestore. The Matadata Parser reads the header sections of the visited web pages and extracts RDF triples from available metadata. Due to its large availability, the preferred metadata format is OpenGraph RDFa, however other formats are allowed as well, as long as they can be converted into RDF. The Data Linker receives the collected triples as input and adds new triples linking visited pages with DBpedia entities. This task is accomplished by both expanding URIs pointed by object properties and by analysing the content of datatype properties such as *tag*, *title*, and *description* with basic NLP techniques in order to find possible matches with DBpedia entries. Finally, the augmented set of triples is processed by a Reasoner module, performing logic entailments in order to classify visited pages. In our prototype the reasoning task is performed by the OWL Lite Reasoner that comes bundled with Apache Jena, but any other OWL reasoner (e.g: Pellet) could fit as well. The result of this process is a semantic user model, built incrementally as the user visits Web pages, in which visited pages are classified and have a hopefully high number of semantic properties linking them each other and to DBpedia. In our prototype system the client part is a standalone application, however, in a production scenario it could be a Web browser plug-in, in order to incrementally build the user profiles as pages are downloaded by the Web browser.

The server part of the system is designed to simulate a content provider scenario and consists in two modules,a *Semantic Recommender*, and a *User Inquirer*, and in a content repository. We assume each content to be addressed towards a specific user stereotype, which is a realistic assumption since many e-commerce companies already do market segmentation analysis. We exploit such knowledge in order to map user characteristics into a specific stereotype and therefore contents to be recommended. More specifically, in our current experimental system we use a decision tree for classifying the user, as shown in Figure 2. Each node is associated with a specific SPARQL query and each arc corresponds to a possible answer to the parent node's query. Stereotypes are identified on the leaves of the tree. When a client connects, it receives the SPARQL query associated with the root node in order to check whether a specific characteristic is present in the user model. The Semantic Recommender module handles the client's answer to the query and fetches content or further queries from the content repository. Due to the hierarchical nature of the decision tree, we expect the number of queries to be asked to the client before being able to identify relevant content to be very small: indeed, in our experimental setting in the worst case six queries were needed.

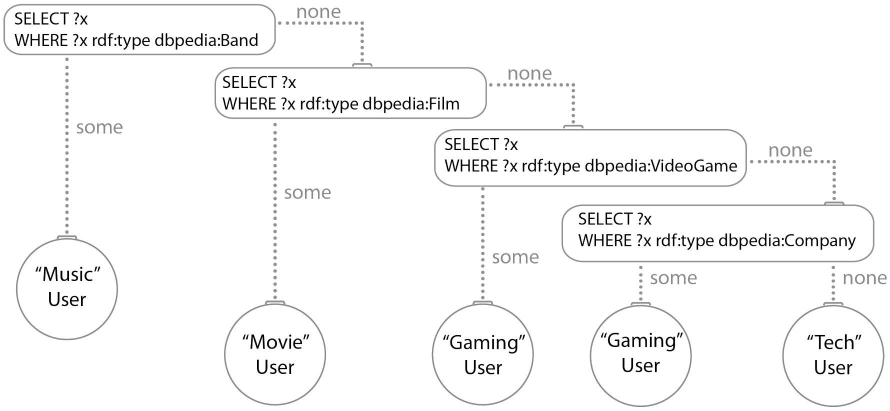


Figure 2. A decision tree with SPARQL queries on the nodes and user stereotypes on the leaves.

4 Structured data augmentation and classification

Metadata commonly embedded in Web pages actually provide a very shallow description of the page's content: the Open Graph protocol itself specifies only four attributes as mandatory (title, image, type, and url) and six object classes (video, music, article, book, profile, and website). However, these informations are a good starting point, especially when a few optional properties too are specified, providing "hooks" to more descriptive ontologies.

Instead of focusing on a particular domain ontology, in this work we have chosen to adopt a general purpose and freely available reference ontology: DBpedia. This choice is motivated by three factors: (i) in a realistic scenario it is impossible to restrict users' Web usage to a particular domain, (ii) authors may describe their contents in ways non compliant to a single taxonomy crafted by a domain expert, therefore, the ontology needs to be the result of a collaborative effort, and (iii) since the modelling task is to be accomplished at client-side, we need an ontology freely accessible by anyone.

The Data Linker module of the system analyses the RDF data extracted from the pages in order to find "hooks" to DBpedia, that are named entities present in DBpedia either linked by an extracted Object Property or present as strings in the body of some Datatype Property. To this aim, properties such as *title* and *tags* are particularly useful since they clearly identify relevant entities. Another interesting property is *description* which contains a very short text summarizing the content of the page: this can easily be processed by means of stopword removal and POS tagging in order to extract all its meaningful substrings that match DBpedia entries. Once these entities have been identified, they are linked to the Web page RDF representation with a *relatedTo* property, defined in the UAC Ontology. All the *rdf:type*, *dc:subjec*, and *db:type* attributes of the linked entity are then imported into the RDF data, in order to provide further information about the contents of the page and to support the classification task.

The classification task is entirely performed by the *Reasoner Module*, which entails the evidence provided by both extracted and augmented statements with class and property axioms provided by the Open Graph specification and by the UAC ontology. The Open Graph specification, as mentioned above, provides six classes, each of them has a unique set of properties: for instance, the OGP "article" class has the "author", "section", and "published_time" properties, so a page including one or more of those properties in its metadata can be easily labelled as a "article" page. On the other hand, the UAC Ontology provides a "relatedTo" property connecting Web pages (classified as "webSite" objects) to DBpedia entries and a "nextInStram" property linking each page to the next one requested by the user. Related DBpedia entries and adjacent (successively and previously visited) pages are then exploited by the reasoner to infer new *rdf:type* attributes for page elements. The result, as shown in Figure 3, is a twofold classification of visited pages, which are labelled according to the entities they are "relatedTo" and to the form of the content, specified by Open Graph metadata.

```
 1  <rdf:Description rdf:about="http://www.sonatype.com/request/2014-
       developer-survey?s2=arba2">
 2     <uac:relatedTo rdf:resource="http://dbpedia.org/resource/
           Open_Source"/>
 3     <uac:relatedTo rdf:resource="http://dbpedia.org/resource/
           Sonatype"/>
 4     <uac:relatedTo rdf:resource="http://dbpedia.org/resource/
           Development"/>
 5     <og:title xml:lang="en">4th Annual Open Source Development
           Survey - Sonatype.com</j.1:title>
 6     <og:locale xml:lang="en">en_US</j.1:locale>
 7     <uac:nextInStream rdf:resource="https://www.surveymonkey.com/s
           /2014_OpenSource?s2=arba2"/>
 8     <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
 9     <og:type xml:lang="en">website</j.1:type>
10  </rdf:Description>
```

Figure 3. A snippet of generated RDF data.

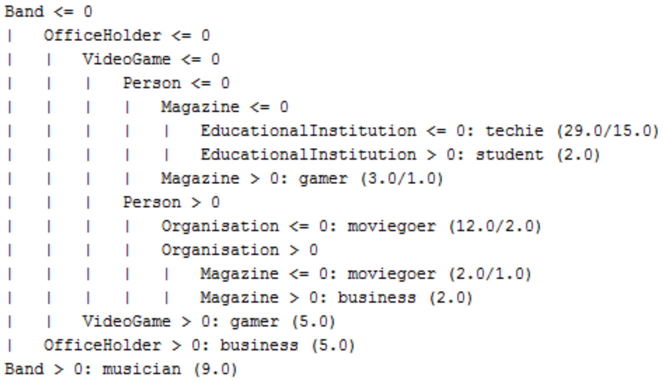
In the example illustrated in 3 it can be noticed how Open Graph properties (lines 5, 9 and 9) are maintained and the original RDF Description element of the Web page is enriched with UAC properties (lines 2, 3, 4, and 7). In this case, the Open Graph property "og:type" (line 9) has the value "website", classifying the item as a Web portal, and the UAC "relatedTo" property has, among others, the "Open_Source" value, which, in DBpedia, is connected to entities such as "Standards" and "Free_Software". By doing so, our model provides a semantic representation of the content visited so far by the user and its form (e.g: website, article, video, ...).

5 Evaluation

Formative tests were performed in order to evaluate the accuracy of the proposed method. In our experiment, we asked a number of volunteers (mostly university students) to let us use their browsing history data, in order to have real-world data. In order to avoid biases, browsing data was asked to be relative to sessions occurred in the five days before the test subjects were asked to supply data, moreover all test subjects were completely unaware of the real purpose of the experiment. After supplying the data, volunteers were asked to review their own browsing history in order to identify different sessions and to point out what they were actually looking for. At the end of a process we were able to identify six user stereotypes, much like market analysts do when performing segmentation analysis. Since we had no real content to provide in this experiment, we only classified users. The six identified stereotypes are: (i) people mostly interested in economics (nicknamed *business*), (ii) mostly interested in courses, seminars, summer schools and other educational events (*student*), (iii) mostly interested in films and tv series (*moviegoer*), (iv) mostly interested in music (*musician*), (v) mostly interested in videogames (*gamer*), and, finally, (vi) people whose main interests are hardware, programming, and technology in general (*techie*). Three iterations of the data gathering and testing process were performed, each time with different volunteers, in order to test our approach with different users with different browsing habits, and different size of the training set. In the first iteration 36 browsing sessions were collected and labelled, in the second 49 and in the third 69.

Over the three iterations, the average number of Web sites visited in a single browsing session was 31.5 and the average number of triples extracted from a browsing session was 472.8.

During each iteration of the evaluation, the *rdf:type* properties of the visited Web pages were considered as features and used to train a Decision Tree algorithm. In this experiment the J48 algorithm [11] was used; in Figure 4 we show an example of a generated tree, built during the third iteration. The nodes of the tree were then replaced with SPARQL queries and then this structure was used to classify a validation set of user models. A ten-fold cross validation approach was used to estimate the accuracy of the system. Table 1 shows the results of the classification over the three iterations of the data set. Our system was compared with the ZeroR predictor, which always returns the mode value of the training set in order to have a baseline. For this formative experiment, only the precision metric (defined as the number of correctly classified instances



```
Band <= 0
|   OfficeHolder <= 0
|   |   VideoGame <= 0
|   |   |   Person <= 0
|   |   |   |   Magazine <= 0
|   |   |   |   |   EducationalInstitution <= 0: techie (29.0/15.0)
|   |   |   |   |   EducationalInstitution > 0: student (2.0)
|   |   |   |   Magazine > 0: gamer (3.0/1.0)
|   |   |   Person > 0
|   |   |   |   Organisation <= 0: moviegoer (12.0/2.0)
|   |   |   |   Organisation > 0
|   |   |   |   |   Magazine <= 0: moviegoer (2.0/1.0)
|   |   |   |   |   Magazine > 0: business (2.0)
|   |   VideoGame > 0: gamer (5.0)
|   OfficeHolder > 0: business (5.0)
Band > 0: musician (9.0)
```

Figure 4. A decision tree built during the third iteration of the experiment

over the total number of instances) was considered. Though precision values are not very high, it is important to point out two limitations of the performed tests: the number of considered browsing sessions is extremely low, due to the fact that only a handful of volunteers let us analyse and use freely their browsing history data; in fact many volunteers dropped out as soon as they realized that their actual browsing history and not some laboratory activity was needed. Secondly, these results were obtained by considering only the rdf:type attribute as feature when building the decision tree. Evaluation and development are ongoing and further experiments, with more test users, more stereotypes, and a richer RDF vocabulary are planned.

Table 1. Average precision of the UAC system and of a ZeroR classifier on the considered data sets.

Data Set size	ZeroR precision	Tree precision
36	0,306	0.639
49	0,195	0.601
69	0,217	0.623

6 Conclusion

In this paper we presented a new approach towards ephemeral personalization on the Web, relying on semantic metadata available on the web and, even though the presented results are still preliminary, the overall outcome is very promising. With the growth of the Web of Data, we expect in the next few years to be able to raise the average number of extracted triples from a browsing session and therefore build more detailed user profiles.

In our opinion this approach could fit particularly well to the application domain of targeted advertising because of three major advantages over the actual cookie-based techniques: (i) our approach can predict whether a user may like a content he/she has never seen before, rather than associate a user with a set of already visited (and potentially disliked) contents (ii) the explicit decision model of the decision tree can easily be reviewed by domain experts, supporting market analysis and knowledge engineering, and (iii) by deploying the user model at client side, the user has total control over his/her own data, addressing many privacy concerns. However, the proposed approach has one major drawback: in order to receive personalized contents, users have to install a client, which may be either a browser plug in or a standalone application. Anyway, this seems to be necessary for providing real privacy and also other works aimed at addressing the privacy issues of online advertising have stated the need of a software agent [6].

References

1. Adida, B.: hgrddl: Bridging microformats and rdfa. Web Semantics: Science, Services and Agents on the World Wide Web 6(1), 54–60 (2008)
2. Bizer, C., Eckert, K., Meusel, R., Mühleisen, H., Schuhmacher, M., Völker, J.: Deployment of rdfa, microdata, and microformats on the web–a quantitative analysis. In: The Semantic Web–ISWC 2013, pp. 17–32. Springer (2013)
3. Cena, F., Likavec, S., Osborne, F.: Propagating user interests in ontology-based user model. AI* IA 2011: Artificial Intelligence Around Man and Beyond pp. 299–311 (2011)
4. Daoud, M., Tamine-Lechani, L., Boughanem, M., Chebaro, B.: A session based personalized search using an ontological user profile. In: Proceedings of the 2009 ACM symposium on Applied Computing. pp. 1732–1736. ACM (2009)
5. Gao, Q., Yan, J., Liu, M.: A semantic approach to recommendation system based on user ontology and spreading activation model. In: Network and Parallel Computing, 2008. NPC 2008. IFIP International Conference on. pp. 488–492 (Oct 2008)
6. Guha, S., Cheng, B., Francis, P.: Privad: practical privacy in online advertising. In: Proceedings of the 8th USENIX conference on Networked systems design and implementation. pp. 13–13. USENIX Association (2011)
7. Lambrecht, A., Tucker, C.: When does retargeting work? information specificity in online advertising. Journal of Marketing Research 50(5), 561–576 (2013)
8. Liu, X.: Generating metadata for cyberlearning resources through information retrieval and meta-search. Journal of the American Society for Information Science and Technology 64(4), 771–786 (2013)
9. Mizzaro, S., Tasso, C.: Ephemeral and persistent personalization in adaptive information access to scholarly publications on the web. In: Proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems. pp. 306–316. AH '02, Springer-Verlag, London, UK, UK (2002), http://dl.acm.org/citation.cfm?id=647458.728228
10. Osborne, F.: A pov-based user model: From learning preferences to learning personal ontologies. In: User Modeling, Adaptation, and Personalization, pp. 376–379. Springer (2013)
11. Quinlan, J.R.: C4. 5: programs for machine learning, vol. 1. Morgan kaufmann (1993)
12. Sieg, A., Mobasher, B., Burke, R.D.: Learning ontology-based user profiles: A semantic approach to personalized web search. IEEE Intelligent Informatics Bulletin 8(1), 7–18 (2007)
13. Tang, J., Leung, H.f., Luo, Q., Chen, D., Gong, J.: Towards ontology learning from folksonomies. In: IJCAI. vol. 9, pp. 2089–2094 (2009)
14. Wood, M.: How facebook is ruining sharing. Weblog post 18 (2011)
15. Yan, J., Liu, N., Ji, L., Hanks, S.J., Xu, Q., Chen, Z.: Indexing semantic user profiles for targeted advertising (Sep 10 2013), uS Patent 8,533,188