Ontology-based Integration of Web Navigation for Dynamic User Profiling

Anett HOPPE, Ana ROXIN, Christophe NICOLLE
CheckSem Research Group,
Laboratoire Electronique, Informatique et Images (LE2I)
Université de Bourgogne, 21068 Dijon, France
{anett.hoppe, ana.roxin, christophe.nicolle}@checksem.fr

The development of technology for handling information on a Big Data-scale is a buzzing topic of current research. Indeed, improved techniques for knowledge discovery are crucial for scientific and economic exploitation of large-scale raw data. In research collaboration with an industrial actor, we explore the applicability of ontology-based knowledge extraction and representation for today's biggest source of large-scale data, the Web. The goal is to develop a profiling application, based on the implicit information that every user leaves while navigating the online, with the goal to identify and model preferences and interests in a detailed user profile. This includes the identification of current tendencies as well as the prediction of possible future interests, as far as they are deducible from the collected browsing information, and integrated expert domain knowledge. The article at hand gives an overview on the current state of the research, the developments made and insights gained.
Keywords: *Semantic Web, Ontologies, SWRL, Big Data reasoning*

1 Introduction

"Big Data" is one of the big buzzwords of our time – culminating in the creation of various congresses and conferences focusing on only that topic during the recent years (e.g. IEEE Congress on Big Data, starting from 2011). The handling of immense amounts of data brings scientists and analysts in a dilemma: On the one hand, using sophisticated analysis techniques might bring best results, but usually come with a higher processing complexity and time that is just not tolerable for most applications. On the other hand, methods known for their efficiency may fail to exploit the data sources in all their depth.

Several research works proposed distinct criteria to define the nature of "Big Data" (e.g. [1]).

The definition largely converges towards the following five:

- volume: massive amounts of data have to be treated,
- velocity: those data arrive in high speed,
- variety: data types and formats are heterogeneous,
- veracity: data are not always sound and have to be verified,
- value: they have an inherent value that has to be discovered by the application.

Applications acting in a Big Data context have to handle all of them in an efficient manner, balancing analysis depth and performance time.

For that very reason, the application of semantic technology is often discarded for a Big Data context. Semantic analysis seems too complex, too costly to be affordable in an environment in which often already very efficient techniques do not come up to the performance necessities. We want to make a case for ontology-based knowledge representation, even when handling vast data amounts. By employing an ontology that has been customised for the application domain to the very detail, the information is limited to those bits and bytes that are actually relevant. Furthermore, we make an effort to avoid performance issues, by decoupling costly analysis steps from the actual, real-time user profiling process (please refer to Section 0 for details).

Furthermore, costly analysis steps have been decoupled from the final system purpose to avoid performance issues.

We demonstrate this approach based on an application in digital advertising. Publishers nowadays have detailed information about their user's navigation behaviour: servers

DOI: 10.12948/issn14531305/19.1.2015.01

capture not only the web pages that were requested by a certain ID, but also the respective time stamps, device information etc. These elements allow insight in usage patterns, but also a deduction of the various contexts, a user might be active in (a distinction between the working environment and private surfing, for example). In the development of our system, we explore the integration of semantic technology to the process, with a close eye on keeping the system in the range of satisfactory performance.

2 Related Work

Traditionally, profiling approaches (following the methodologies applied in document indexing) use a keyword-based representation to summarise source documents and user interests in an economical way. The limitedness of this data structure raises problems when treating natural language. Synonyms, ambiguities and simple spelling errors cause the system to discover relations where there are none (e.g. when encountering synonyms) or to not discover those that are imminent (in the case of homonyms or spelling errors).

First attempts to alleviate this shortcoming explored the construction of semantic networks. Starting from a base of fixed key-terms, semantic relations between terms were estimated based on co-occurrence in the document base and, new terms added if a relationship was judged likely [2]. The main focus of the researchers was to tackle the problem of homonymy, that refers to terms in natural language that are written the same, but refer to different semantic concepts depending on their context of usage [3]. However, even the combination of elaborate algorithms based on semantic networks did not finally solve the issue [4], at least not without adopting external knowledge sources. For example in [4], WordNet [5] has been used to create links to semantically defined entities.

As a result, more and more researchers explore the usefulness of ontologies for profiling purposes, e.g. [3], [6], [7]. The integration points vary from the usage of ontologies as a background knowledge repository to the usage of ontology-shaped profile constructions. Numerous approaches use structured open linked data [8] as domain knowledge. A lot of them refer to WordNet [9–11], DBPedia [12] or the Open Directory Project (ODP) [13]. Those resources differ in the degree of structure that they induce to the contained concepts. WordNet's relational structure, for example, was obtained by manually grouping terms into "synsets", words that bear a synonymous relationship. The ODP is a community-driven project to classify web content into a given set of categories. There are several works that refer to the ODP as an ontology. However, even though the set of possible relationships extends the mere taxonomic ones, those are not widely used – in consequence, ODP could, at most, be considered a light-weight ontology [14]. Above methods make use from structured knowledge resources, but are far from exhausting their potential.

In turn, there are two recent approaches that make use of full-fledged ontologies for reference: [3] proposed a ontology-based profiling system that relies on the Yago ontology [15]. In [7], the researchers assume a not specified domain ontology as a base for their abstract system.

Finally, the chosen knowledge structure may be used in different ways. First techniques aimed to maintain the keyword-based representation, but use the ontologies for specification and alteration of the keyword space. On the one hand, the analysis of the whole keyword set allowed disambiguating some of them. On the other hand, "synsets" or similar relations can be used to extend the keywords space by synonymous or closely related terms [16]. Chua et al. [17] use them to group semantically close terms to clusters and reduce the feature space for more efficiency.

More recent propositions advance with respect to these approaches by using an ontological structure for the profile itself. This enables them to not only extract additional concepts from the knowledge resource, but also use the relationships to enhance the user's characterization. Calegari et al. [3] ex-

DOI: 10.12948/issn14531305/19.1.2015.01

tract parts of the Yago ontology that are related to the terms that were found in the documents of the user, the choice of concepts is based on the Spreading Activation algorithm [18].

From the above represented approaches, especially the most recent ones [3], [7] show similarities to our vision. However, they both rely on keyword-based representations for the characterisation of manifold resources of the users, that, when included to the user profile, are enriched with semantics.

3 System Design

The final goal of our research and development is a system that integrates with numerous, contradicting demands: The application needs it to make reliable deductions at runtime, which explore the depth of the input data to the maximum. The considerate design of the central data structure has thus an important significance. The result of the design process that has been accompanied by domain experts is presented in the following sections. However, the limitation of the analysis focus to highly relevant concepts alone does not bring the system to comply with the standards of Big Data. Hence, we also adapted the system design to avoid performance losses due to costly working steps. A description of the taken adaptations will follow the description of the ontology in Section 0.

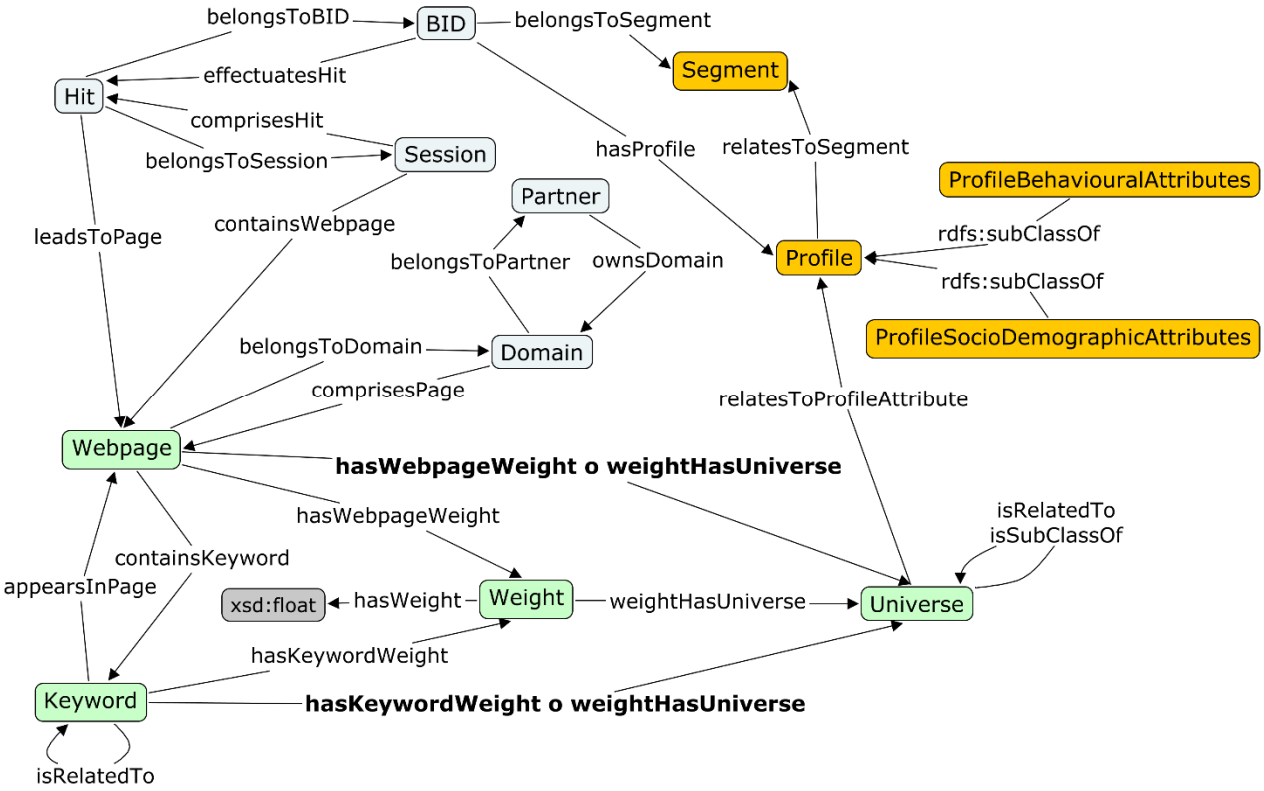

Fig. 1. Schematic view - upper-level concepts of the ontology

3.1 The Ontology

The design principles that guided the conception of the ontology are straight-forward:

- include all concepts that are relevant to the profiling process;
- limit the conception to the essential;
- adapt to the application domain to minimise the complexity of the data structure;
- keep the design modular in the customised parts to allow transfer to other domains.

The result is a data structure that bears highly generic parts (e.g. a user identification string, certain widely used profile attributes), but also modules that tailored to fit the needs of digital advertising (e.g. the topic categoriza-

DOI: 10.12948/issn14531305/19.1.2015.01

tion scheme).
We engage in a closer look on the included entities in the following paragraphs. For a graphical overview, 0 shows a schema of the upper-level classes in the conceived ontology. In light grey, the entities that contribute the navigational history of the user, in green the ones that capture the semantic information about each web resource, in yellow the elements that constitute the final user profile. Please note that the figure is limited to the object properties connecting the concepts among each other. Each of the concepts has several data type properties attached which describes it in more detail. These data type properties will be mentioned in the later passages that take a closer look at the included concepts and their functions.

BID. The "`BID`" stands for "browser identification" and is the central unit in the ontology. Included in every cookie, a `BID` identifies a user whenever he re-visits the website. However, in detail this functioning relates a certain browser on a certain machine, therefore the name "browser-" and not "user ID". The `BID` is used to group the log entries that belong to a single ID and to assess the sequence of pages that is visited. In the ontology, the `BID` is the class that is connected (a) to the effectuated hits, (b) to the visited web pages, (c) to the classes defining the profile, (d) to the segment classes that apply for the respective user. 0 shows the OWL definition of the concept `BID` and its related data type properties, in Turtle syntax.

```
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix mm_base: <http://www.checksem.fr/MindMinings/mm_base#> .
@base <http://www.checksem.fr/MindMinings/mm_base> .
<http://www.checksem.fr/MindMinings/mm_base> rdf:type owl:Ontology .
###  http://www.checksem.fr/MindMinings/mm_base#BID
:BID rdf:type owl:Class ;
     rdfs:comment "BID: a BID is identifying a browser based on a cookie."@en .
###  http://www.checksem.fr/MindMinings/mm_base#hasBID
:hasBID rdf:type owl:DatatypeProperty ,
                 owl:FunctionalProperty ;
        rdfs:domain :BID ;
        rdfs:range xsd:string .
```

Listing 1. Definition and datatype properties for the class "BID" (Turtle)

3.1.1 Context Entities

Some of the entities within the ontology are defined by the commercial ecosystem. The company concludes a contract with online publishers to provide enhanced analysis of their usage logs. Thus, the content that has to be included in the analysis process is determined by the partner, the domains he owns and the web pages that are reachable below this domain. Each partner has different amounts information appertaining to him, possibly extended by collaborations with other actors on the market. Hence, information about the partner and his possible coalitions are crucial to determine which facts to include in the analysis process. Logs are grouped by partner, to avoid any information leakage. We thus included the following entities for the context:

Partner: The "`Partner`" is a society that has signed a contract for the treatment of their data. Each partner is identified by a `WID`, "web identification". All domain and web pages that belong to a partner will have this ID attached. 0 shows the OWL definition of this concept along with its data type properties (in Turtle syntax, shortened, as the prefix definition as in 0).

DOI: 10.12948/issn14531305/19.1.2015.01

```
###  http://www.checksem.fr/MindMinings/mm_base#Partner
:Partner rdf:type owl:Class ;
         rdfs:comment "WID: a WID designates a client of the           advertising
service provider.  Each WID can own one
         or more domains on which the service provider collects
         data."@en .
###  http://www.checksem.fr/MindMinings/mm_base#hasName
:hasName rdf:type owl:DatatypeProperty ;
         rdfs:domain :Partner ;
         rdfs:range xsd:string .

###  http://www.checksem.fr/MindMinings/mm_base#hasWID
:hasWID rdf:type owl:DatatypeProperty ,
                 owl:FunctionalProperty ;
        rdfs:domain :Partner ;
        rdfs:range xsd:string .
```

Listing 2. Definition and datatype properties for the class "Partner" (Turtle)

Domain: Referring to the official Domain Name System (DNS) [19], the domain in the project context means the string that results from the combination of second-level domain and top-level domain. All web pages and sub-domains subordinated to the domain will be related to it. For example: the URL "`http://lentreprise.lexpress.fr/index.html`" refers to the domain "`lexpress`", "`lentreprise`" is the respective sub-domain, "`index.html`" identifies the specific page to display. The definition of this concept is given below (see 0).

```
###  http://www.checksem.fr/MindMinings/mm_base#Domain
:Domain rdf:type owl:Class ;
        rdfs:comment """Domain: a domain is described by the following pseudo-code:
        IF URL == \"http://lentreprise.lexpress.fr/index.html\"
        DOMAIN_URL = \" lexpress.fr\"
        ENDIF"""@en .
###  http://www.checksem.fr/MindMinings/mm_base#hasURL
:hasURL rdf:type owl:DatatypeProperty ,
                owl:FunctionalProperty ;
        rdfs:range xsd:string ;
        rdfs:domain [ rdf:type owl:Class ;
                      owl:unionOf ( :Domain
                                    :Hit
                                    :Webpage
                                  )
                    ] .
```

Listing 3. Definition and datatype properties for the class "Domain" (Turtle)

3.1.2 Data Entities

The entities in this section stem from internal data treatment of the collaboration partner. For that, they provide additional contextual information to the web pages extracted from the navigation logs. However, their origin are basic analytic steps (as computing the duration of a page view from the time stamps accompanying it, for instance), as opposed information deduced from enhanced statistics or the application of machine learning techniques.

Hit: The "`Hit`" comprises all information about a single user action. That is, whenever a page is requested from the server, this is logged as one hit. Included in the class is all information related to that entity -- the time stamp, the requested URL etc. Thus, it capsules all information that can be related to a single event connected to a user. Via relations, the "`Hit`" connects the user with a set of web pages; several hits are grouped in a session (see 0).

DOI: 10.12948/issn14531305/19.1.2015.01

```
###  http://www.checksem.fr/MindMinings/mm_base#Hit
:Hit rdf:type owl:Class ;
     rdfs:comment """Hit: the hit is the basic entity collected on WID domains. It
corresponds to an event on a WID domain, and has some
     native variables which characterize it such as its timestamp, its
     URL, its user agent and the referer (the page visited just
     before)."""@en .
###  http://www.checksem.fr/MindMinings/mm_base#isClick
:isClick rdf:type owl:DatatypeProperty ,
                  owl:FunctionalProperty ;
         rdfs:comment "Click: a click is an event made by a web user; denotes the
action of a web user on a banner." ;
         rdfs:domain :Hit ;
         rdfs:range xsd:boolean .

###  http://www.checksem.fr/MindMinings/mm_base#hasURL
:hasURL rdf:type owl:DatatypeProperty ,
                 owl:FunctionalProperty ;
        rdfs:range xsd:string ;
        rdfs:domain [ rdf:type owl:Class ;
                      owl:unionOf ( :Domain
                                    :Hit
                                    :Webpage
                                  )
                    ] .

###  http://www.checksem.fr/MindMinings/mm_base#hasTimestamp
:hasTimestamp rdf:type owl:DatatypeProperty ,
                       owl:FunctionalProperty ;
              rdfs:range xsd:integer ;
              rdfs:domain [ rdf:type owl:Class ;
                            owl:unionOf ( :Hit
                                          :Session
                                        )
                          ] .
```

Listing 4. Definition and datatype properties for the class "Hit" (Turtle)

Session: A "Session" is a sequence of hits, grouped by the fact that the distance between the time stamp of one page view and the subsequent page does not exceed thirty minutes. The class definition and related datatype properties in Turtle syntax can be found in 0.

```
###  http://www.checksem.fr/MindMinings/mm_base#Session
:Session rdf:type owl:Class ;
         rdfs:comment "Session: two hits of a BID belong to the same session if the
difference of their time-stamps does not exceed 30 minutes."@en .

###  http://www.checksem.fr/MindMinings/mm_base#hasDuration
:hasDuration rdf:type owl:DatatypeProperty ,
                      owl:FunctionalProperty ;
             rdfs:domain :Session ;
             rdfs:range xsd:integer .
###  http://www.checksem.fr/MindMinings/mm_base#hasTimestamp
:hasTimestamp rdf:type owl:DatatypeProperty ,
                       owl:FunctionalProperty ;
              rdfs:range xsd:integer ;
              rdfs:domain [ rdf:type owl:Class ;
                            owl:unionOf ( :Hit
                                          :Session
                                        )
                          ] .
```

Listing 5. Class definition and datatype properties for the class "Session" (Turtle)

DOI: 10.12948/issn14531305/19.1.2015.01

3.1.3 Analysis Entities

The content analysis process demands the integration of the following classes (green-coloured in **Error! Reference source not found.**):

Keyword: A "Keyword" is a term that describes one concept contained in a web page. The Keyword class will be used to handle their disambiguation using external knowledge sources, by means of external URIs that link to external knowledge sources such as DBPedia and WordNet. The instances of the "Keyword" class are related with the web pages that they appear in and the universes that they are semantically related to. Furthermore, both of these relations may be attributed with a weight, measuring the degree of membership.

```
###  http://www.checksem.fr/MindMinings/mm_base#Keyword
:Keyword rdf:type owl:Class .

###  http://www.checksem.fr/MindMinings/mm_base#hasURI
:hasURI rdf:type owl:DatatypeProperty ;
        rdfs:domain :Keyword ;
        rdfs:range rdf:XMLLiteral .
```

Listing 6. Definition and datatype properties for the class "Keyword" (Turtle)

Webpage: The class "Webpage" envelops all content pages that are extracted for a certain (sub-) domain. Every parsed web page will constitute an individual within the ontological structure and relate with other entities that qualify its content. It is identified by its Unified Resource Location (URL), which is included in a data property connected to it. A web page belongs to a certain domain and is thus connected to such with an appropriate relation. Furthermore is related to a user by a hit that he effectuated. During the analysis process, a web page is additionally related to a set of pertinent keywords and, respectively to the universes that it covers. A condensed listing of the OWL source describing the class "Webpage" can be found in 0.

```
###  http://www.checksem.fr/MindMinings/mm_base#Webpage
:Webpage rdf:type owl:Class .

###  http://www.checksem.fr/MindMinings/mm_base#hasURL
:hasURL rdf:type owl:DatatypeProperty ,
                 owl:FunctionalProperty ;
        rdfs:range xsd:string ;
        rdfs:domain [ rdf:type owl:Class ;
                      owl:unionOf ( :Domain
                                    :Hit
                                    :Webpage
                                  )
                    ] .
```

Listing 7. Definition and datatype properties for the class "Webpage" (Turtle)

Universe: The term "Universe" refers to a certain content category and the keywords that are related to it. Thus, every Universe will carry the name of the category it depicts, and bear close relations to the reference keywords that are associated with the respective content domain.

```
###  http://www.checksem.fr/MindMinings/mm_base#Universe
:Universe rdf:type owl:Class .

###  http://www.checksem.fr/MindMinings/mm_base#Universe_Actualities
:Universe_Actualities rdf:type owl:Class ;
                      rdfs:subClassOf :Universe .

###  http://www.checksem.fr/MindMinings/mm_base#Universe_Sports
```

DOI: 10.12948/issn14531305/19.1.2015.01

```
:Universe_Sports rdf:type owl:Class ;
                 rdfs:subClassOf :Universe .

###  http://www.checksem.fr/MindMinings/mm_base#Universe_Travel
:Universe_Travel rdf:type owl:Class ;
                 rdfs:subClassOf :Universe .


###[more sub-classes exist but have been omitted for clarity]
```

Listing 8. Class definition and data type properties for the class "Universe" (Turtle)

3.1.4 Profile Entities

The final goal of all computing efforts is to build a semantically enhanced profile representation for every considered user.

Profile: "`Profile`" is the main class containing the attributes that define the user profile (0 gives the complete definition of the concept). This comprises the elements stemming from the content analysis of the web pages, by linking it with the universes that were discovered therein; but also attributes that may be deduced from those content attributes. In consequence the `Profile` class contains two sub-classes that group the elements into socio-demographic attributes (such as age, location etc.) and behavioural attributes (such as the browser used or the affinity to certain brands).We chose to divide each of those sub-classes into a number, as to signify commercially interesting divisions of the attributes. For example, the value of the property "`hasAge`" is defined by choosing to link a profile with one or more of the individuals "`Age 15-24`", "`Age 25-34`", "`Age 35-49`", "`Age 50-64`", "`Age above65`", "`Age Child`", "`Age Pre-Teenager`", or "`Age Teenager`".

```
###  http://www.checksem.fr/MindMinings/mm_base#Profile
:Profile rdf:type owl:Class .

###  http://www.checksem.fr/MindMinings/mm_base#Socio-Demographic_Attribute
:Socio-Demographic_Attribute rdf:type owl:Class ;
                             rdfs:subClassOf :Profile .

###  http://www.checksem.fr/MindMinings/mm_base#Behavioural_Attribute
:Behavioural_Attribute rdf:type owl:Class ;
                       rdfs:subClassOf :Profile .


###[more sub-classes exist but have been omitted for clarity]
```

Listing 9. Definition and data type properties for the class "Profile" (Turtle)

Segment: One of the key features of the ontology lies in its capability to automatically infer the attribution of an individual to a certain segment (0). Using the attribution of a user individual to certain of the above described tranches, more complex notions can be specified. The segment class captures exactly those more complex profile entities that may be constructed using profile features ("a female person living in a household with children belongs to the segment "mother"), content features ("a person reading 90% of the times on pages that treat sports-related topics is a sports-fan") or a combination of both. The individuals assigned to a class of type "`Segment`" are those that comply with the constraints or rules that were imposed to define the segment[1].

[1] Please note that this and all following examples have been chosen deliberately simple for the sake of clarity. Realistic segments are most likely to be more complex.

DOI: 10.12948/issn14531305/19.1.2015.01

```
.
###  http://www.checksem.fr/MindMinings/mm_base#Segment
:Segment rdf:type owl:Class ;
         rdfs:comment "Segment: a segment is commercial word used to designate a set
of BIDs satisfying some rules. (In Data Mining
         rather called "cluster", yet, segment is commonly used to
         facilitate communication with clients."@en .

###  http://www.checksem.fr/MindMinings/mm_base#Seg_Parent
:Seg_Parent rdf:type owl:Class ;
            owl:equivalentClass [ rdf:type owl:Class ;
                                  owl:intersectionOf ( :Family_Child
                                      [ rdf:type owl:Class ;
                                        owl:unionOf ( [ rdf:type owl:Restriction ;
                                                        owl:onProperty :hasAge ;
                                                        owl:hasValue :Age_25-34
                                                      ]
                                                    [ rdf:type owl:Restriction ;
                                                      owl:onProperty :hasAge ;
                                                      owl:hasValue :Age_35-49
                                                    ]
                                                    [ rdf:type owl:Restriction ;
                                                      owl:onProperty :hasAge ;
                                                      owl:hasValue :Age_50-64
                                                    ]
                                                  )
                                      ]
                                  )
                                ] ;
            rdfs:subClassOf :Segment ;
            rdfs:comment    "Segment    defined    by    child    presence    in    the
household and the estimated age of the BID"@en .
###  http://www.checksem.fr/MindMinings/mm_base#Seg_Mother
:Seg_Mother rdf:type owl:Class ;
            owl:equivalentClass[rdf:type owl:Class ;
                                owl:intersectionOf (:Seg_Parent
                                                    [rdf:type owl:Class ;
                                                     owl:complementOf :Seg_Father
                                                       ]
                                                    )
                                ] ;
            rdfs:subClassOf :Seg_Parent .
```

Listing 10. Definition and datatype properties for the class "Segment" (Turtle)

3.1.5 Constructional Entities

In order to allow weighting relations between certain entities, an additional class was added to the ontology. The concept "`Weight`" may be used to specify a numerical value of membership to a certain property, as for example the relation between a web page and a universe. A web page may cover a set of various topics, each of them to a certain degree. The class `Weight` enables us to model this fact within our ontology.

The individuals from the class "`Weight`" (0) carry a data type property containing the numerical value that quantifies the weight of the relation (namely the "`hasWeightValue`" property). The relation between the web page and the universe, named "`hasUniverse`" at the moment, is then specified as being a composition of two other relations: "`hasWebpageWeight`", relating the web page with the weight concept and "`weightHasUniverse`" that then concludes the relation to the respective universe (see 0 and 0).

DOI: 10.12948/issn14531305/19.1.2015.01

```
###  http://www.checksem.fr/MindMinings/mm_base#hasKeywordWeight
:hasKeywordWeight rdf:type owl:ObjectProperty ;
                  rdfs:domain :Keyword ;
                  rdfs:range :Weight .

###  http://www.checksem.fr/MindMinings/mm_base#weightHasUniverse
:weightHasUniverse rdf:type owl:ObjectProperty ;
                   rdfs:range :Universe ;
                   rdfs:domain :Weight .

###  http://www.checksem.fr/MindMinings/mm_base#hasUniverse
:hasUniverse rdf:type owl:ObjectProperty ;
             rdfs:range :Universe ;
             owl:inverseOf :belongsToUniverse ;
             owl:propertyChainAxiom ( :hasKeywordWeight
                                      :weightHasUniverse
                                    ) .
```

Listing 11. Usage of "Weight" in example object property (Turtle)

The same has been done for the relation between a keyword and a universe (quantifying how much a keyword is actually associated to a certain category), between a profile and a universe (quantifying how much importance the universe in question has for the description of the profile).

As such, the "`Weight`" concept allows us to put a measurement of importance on some of the relations within the ontology. In consequence, we are able to not only express binary relations ("mother AND some web pages that talk about sports" means "`SportyMom`"), but insert a new level of expressiveness by allowing quantification: "to a certainty of 0.8 a mother AND more than 90% of pages treat topics related to sports" means "`SportyMom`". This enhanced expressiveness extends relationships whenever it seems useful. Apart from above example this includes, for instance, the relationship between a topic universe and a web page (quantifying the importance of a certain topic for the message of the resource's content) or the relationships among "`Keywords`" (quantifying the degree of semantic relatedness between them).

```
###  http://www.checksem.fr/MindMinings/mm_base#Weight
:Weight rdf:type owl:Class ;
        rdfs:comment "Hub entity to attach a numerical value to a relationship,
quantifying the degree to which it applies"@en .

###  http://www.checksem.fr/MindMinings/mm_base#hasWeightValue
:hasWeightValue rdf:type owl:DatatypeProperty ;
                rdfs:comment   "relating   the   Weight   concept   with   its
numerical value "@en .
```

Listing 12. Definition and datatype properties for the class "Weight" (Turtle)

The above-described ontology comes to use in a completely automatic, web-based system for the treatment of web resources and user data. The web resources are monitored and qualified continuously, decoupled from the user profiling process. This procedure enables to keep up with huge amounts of data and to have the content information necessary for the profile construction ready whenever a user surfs an indexed page. The final user profile is constructed using the information obtained from the indexation process, information that can be directly retrieved from the cookie (e.g. the user agent) and the knowledge that stems from the company's internal evaluation processes.

DOI: 10.12948/issn14531305/19.1.2015.01

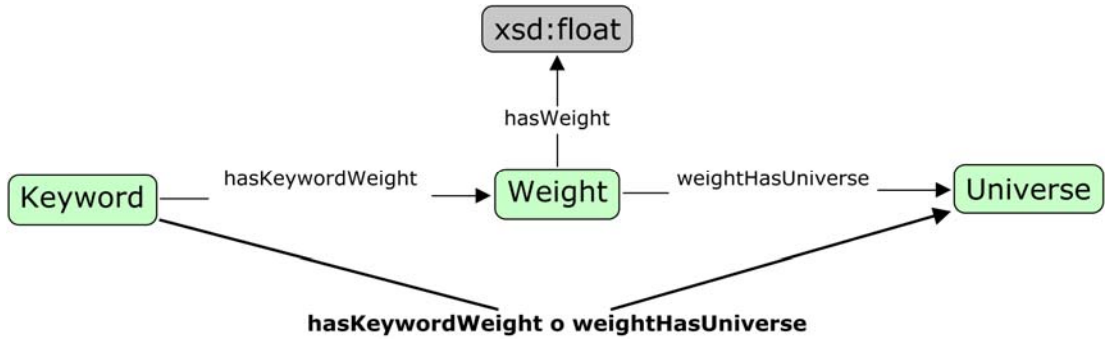

Fig. 2. Integration of the "Weight" concept in the MindMinings ontology

3.2 System Architecture

The main focus of the article at hand has been the description of the conceived structure for knowledge representation within the profiling system. For the sake of completeness, we want to give a short overview on the design of the surrounding profiling/analysis architecture.

Enhanced syntactic and semantic analysis can be computationally expensive. For this reason they have been avoided when handling vast amounts of data up to this point. In an industrial context, the profiling application will have to handle a multitude of data instances that may arrive in high speed. One of our collaboration partners specifies the number of arriving user events to at least 150 million per month, for a single publisher site. Hence, in a simplistic calculation, the system will have to treat about 60 events per second for each single client. (Of course, user activity is not evenly distributed throughout the day. Periods of higher activity, such as the early evening, will account for a huge percentage of user events.)

Performing semantic analysis at runtime might cause the system delays that are not acceptable for the application context. Hence, we decoupled the semantic analyses from actual user activity. In doing so, we benefit from the practical setting in the industrial context: Due to privacy concerns, every online publisher has only access to those parts of the navigation logs that happen on her websites or those of collaborators. Even though this might involve a considerable amount of web resources, it is still a limited set of contents. Those can be continuously monitored and analysed, the relevant semantic information be kept in the system. In consequence, the actual profiling task that is performed at runtime is reduced to the connection of the already available semantic page information according to each user's individual behaviour, and the deduction of inherent patterns.

0 shows an overview on the high-level building blocks and work-flows. On the left hand side, the web pages ("WP") enter the asynchronous analysis process. The results of their semantic qualification are directly added to the ontology. On user activity, that information are related with a user ID, user agent and session information and ontological inference is used to deduce the relevant customer segments.

The semantic page information within the system will be updated on a regular basis, based on the lifecycle of the indexed pages. To preserve a maintainable knowledge base, contents that are vital have to be identified; contents that are outdated or uninteresting for the user base have to be discarded. The concrete measurement for site vitality is still to be developed and tested. However, incoming and outgoing links (and the vitality of the pages they lead to), the age of the page and the reappearance of its core concepts in novel resources seem to be a good starting point for a pertinent, automatic measurement.

DOI: 10.12948/issn14531305/19.1.2015.01

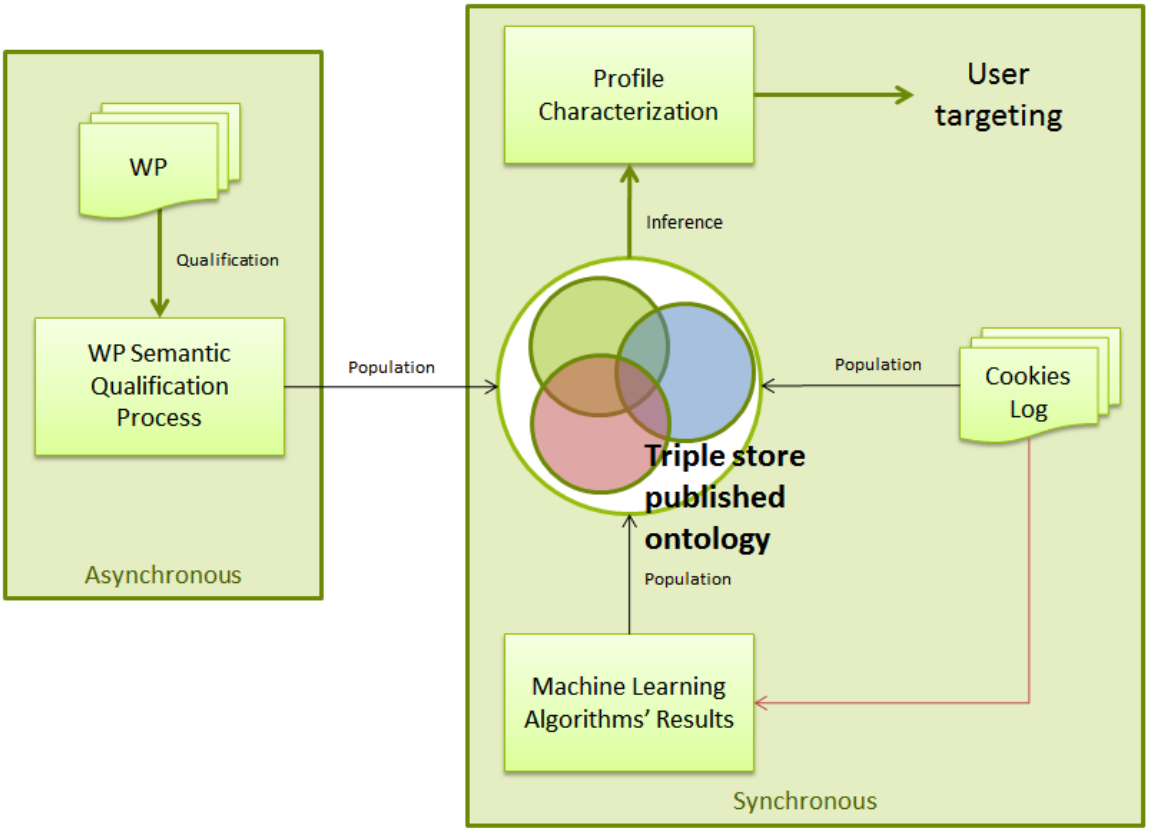

Fig. 3. Overview of the work ow within the MindMinings profiling system
("WP" means Web page)

4 Conclusion and Future Work

The core focus of the article at hand is the presentation of a data structure that has been conceived to support user profiling in digital advertising. The ontology represents expert knowledge about the entities, their relationships and surrounding information flows that are essential to the profiling process.

The ontology is integrated in a system that performs semantic analysis based on the structured input files that contain the navigation history of a multitude of users. Critical design decisions have been taken in consideration of the specific profiling needs of digital advertising and the pragmatics of the final application context. The system will have to perform in a Big Data environment and has thus been subject to focussed adaptations with respect to the before cited criteria of Big Data contexts:

Volume: In a realistic working environment, the system will have to cope with vast amounts of entering data, stemming from a multitude of commercial partners. A single publisher site accounts for about 150 million user events per month in average – and all those events have to be processed and integrated. To respond to this issue, the underlying ontology has been designed to capture only the information relevant to the profiling task, discarding additional available information that has no value for the specific application domain.

Velocity: Semantic analysis is often considered too costly to be applied in a Big Data environment. To avoid it to delay the process, the time-consuming information has been decoupled from the actual profiling on run-time. Web documents are monitored and analysed in the system when they appear online. When the actual user activity happens, only the respective relationships have to be added to the ontological structure. Moreover, the used RDF database [20] allows the population of the ontology at very high speed (about 35000 triples per second). The inference processes necessary for the profile construction are performed in quasi

DOI: 10.12948/issn14531305/19.1.2015.01

real-time.

Variety: A publishing site usually features a multitude of different document types and format. For the time being, we focus on the analysis of textual documents, tackling the difficulties that emerge from various types of html-formatting through different websites. The system is designed highly modular to allow later extension by further information sources. One could imagine additional modules for the semantic annotation of images or videos that feed their information to the same central repository. However, we chose to focus on one main information source for the deployment of the initial system, to allow fast validation of the presented approach.

Veracity: Ontologies have their strengths in the validation of information and the discovery of contrasting statements. The reason functionality is specifically designed to test all possible deductions, discover contradictions, and to propose strategies for their dissolution. Furthermore, the explicit interconnection with external knowledge repositories allows leveraging their contents for this purpose. Each concept within the ontology carries a unique identifier (a URI) that is mapped to its counterpart in an external knowledge resource by means of predicates such as "`owl:sameAs`". This enables to evaluate the information obtained from the analysis of a document for its meaningfulness. Links and statements that seem too far off the information in the knowledge repository will be discarded from the profile; additional relationships from the external repository added as necessary.

Value: The data we are working on are the navigation logs of actual users as they are already used to determine interests for content recommendation. Several publications outline methods for their statistical analysis and evaluate the gain in user involvement (e.g. [21])

In the close future, we will engage in detailed performance testing, preferable in direct comparison to a similar profiling system in industrial usage. First tests have been realised in a laboratory environment and turned out promising, with an average response time well below one second. Those, however, have to be verified in a more realistic setting. For the moment, the integration of external resources is limited to one, namely DBPedia. Currently, we are experimenting with alternative connectors to repositories such as Yago, Freebase and Wolfram Alpha [22]. These references differ in their structuring of the knowledge, their query language, the frequency of updates and, importantly, response time. The goal will be to find a good balance between the extensiveness of the accessible knowledge and the time that has to be invested to extract it. Furthermore, when relying on diverse information sources for our analyses, we will have to decide how to balance their influences on the final relationship structure in the dynamic ontology and what theoretical framework to use for their aggregation.

References

[1] M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. B. J. Manyika, “Big Data: The next frontier for innovation, competition, and productivity,” McKinsey Global Institute, 2011.

[2] G. Gentili, A. Micarelli, and F. Sciarrone, “Infoweb: An adaptive information filtering system for the cultural heritage domain,” Applied Artificial Intelligence, vol. 17, no. 8–9, pp. 715–744, 2003.

[3] S. Calegari and G. Pasi, “Definition of User Profiles Based on the YAGO Ontology.,” in IIR, 2011.

[4] B. Magnini and C. Strapparava, “Improving user modelling with content-based techniques,” in User Modeling 2001, Springer, 2001, pp. 74–83.

[5] G. A. Miller, “WordNet: a lexical database for English,” Communications of the ACM, vol. 38, no. 11, pp. 39–41, 1995.

[6] M. Speretta and S. Gauch, “Miology: A web application for organizing personal domain ontologies,” in Information, Process, and Knowledge Management, 2009. eKNOW’09. International Conference on, 2009, pp. 159–161.

[7] Z. Su, J. Yan, H. Ling, and H. Chen,

DOI: 10.12948/issn14531305/19.1.2015.01

"Research on personalized recommendation algorithm based on ontological user interest model," Journal of Computational Information Systems, vol. 8, no. 1, pp. 169–181, 2012.

[8] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," International journal on semantic web and information systems, vol. 5, no. 3, pp. 1–22, 2009.

[9] E. Agirre and G. Rigau, "Word sense disambiguation using conceptual density," in Proceedings of the 16th conference on Computational linguistics-Volume 1, 1996, pp. 16–22.

[10] S. Banerjee and T. Pedersen, "An adapted Lesk algorithm for word sense disambiguation using WordNet," in Computational linguistics and intelligent text processing, Springer, 2002, pp. 136–145.

[11] L. Dey, S. Singh, R. Rai, and S. Gupta, "Ontology aided query expansion for retrieving relevant texts," in Advances in Web Intelligence, Springer, 2005, pp. 126–132.

[12] G. Demartini, C. S. Firan, and T. Iofciu, "L3s at inex 2007: Query expansion for entity ranking using a highly accurate ontology," in Focused Access to XML Documents, Springer, 2008, pp. 252–263.

[13] E. Gabrilovich and S. Markovitch, "Feature generation for text categorization using world knowledge," in IJCAI, 2005, vol. 5, pp. 1048–1053.

[14] R. Mizoguchi, "Part 3: Advanced course of ontological engineering," New Generation Computing, vol. 22, no. 2, pp. 193–220, 2004.

[15] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in Proceedings of the 16th international conference on World Wide Web, 2007, pp. 697–706.

[16] J. Bhogal, A. Macfarlane, and P. Smith, "A review of ontology based query expansion," Information processing & management, vol. 43, no. 4, pp. 866–886, 2007.

[17] S. Chua and N. Kulathuramaiyer, "Semantic feature selection using WordNet," in Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence, 2004, pp. 166–172.

[18] A. Katifori, C. Vassilakis, and A. Dix, "Ontologies and the brain: Using spreading activation through ontologies to support personal interaction," Cognitive Systems Research, vol. 11, no. 1, pp. 25–41, 2010.

[19] P. Mockapetris and K. J. Dunlap, Development of the domain name system, vol. 18, no. 4. ACM, 1988.

[20] "Stardog - A Developer's Best Friend." 2013.

[21] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen, "Demographic prediction based on user's browsing behavior," in Proceedings of the 16th international conference on World Wide Web, 2007, pp. 151–160.

[22] B. Walters, "Wolfram| Alpha, A New Kind of Science," 2011.

Anett HOPPE received a Master's degree in Data and Knowledge Engineering from the University of Magdeburg, Germany and is currently a PhD student at the Checksem research group, part of the LE2I (Laboratory of Electronics, Informatics and Images), UMR CNRS 6306 at the University of Burgundy, Dijon, France. Her current research focus is the usage of Semantic Web technologies for applications in business intelligence with a main focus on the development of profiling solutions for personalised Web applications. Further interests are knowledge engineering, natural language analysis and data analysis.

DOI: 10.12948/issn14531305/19.1.2015.01

Ana ROXIN is an associate professor at the department of Informatics, Electronics and Mechanics (IEM) from the University of Burgundy. She is a member of the Checksem research team, part of the LE2I (Laboratory of Informatics, Image and Electronics) UMR CNRS 6306. She was involved in several national and European projects (e.g. EU FP7 ASSET, TELEFOT) addressing information recommendation to the user. Her main research interest concern: Semantic Web technologies, knowledge engineering, data interoperability and user profiling in a Big Data context.

Christophe NICOLLE is professor in the Computer Science department at the University of Burgundy. He received his Ph.D. in Computer Science in 1996. Since, he is a member of the LE2I laboratory (Electronics, Informatics and Image) at the University of Burgundy. His research interests include interoperability of heterogeneous information systems and the optimization of process and resources using semantics, combinatory and logical rules. Since 2001, he works on the Active3D project dedicated to the development of a semantic framework for the management of buildings during their life-cycle. In 2005, he participated in the creation of the Active3D Company. The company develops a web collaborative platform for facility management. Currently, the Active3D platform manages more than 60 millions of square meters of buildings.

DOI: 10.12948/issn14531305/19.1.2015.01