Undefined 0 (0) 1
IOS Press

Ontology-based User Profile Learning from heterogeneous Web Resources in a Big Data Context

Hoppe Anett [a], Roxin Ana [a], Nicolle Christophe [a]
[a] *CheckSem Research Group, Laboratoire Electronique, Informatique et Image (LE2I), Université de Bourgogne, BP 47870, 21078 Dijon, France, E-mail: firstname.lastname@checksem.fr*

Abstract. With the emergence of real-time distribution of online advertising space ("real-time bidding"), user profiling from traces left by online navigation reaches a new importance. The ability to distinguish user interests based on implicit information as it is contained in navigation logs, enables online advertisers to target customers without interfering with their activities. Current techniques apply traditional methods as statistics and machine learning, but also suffer from their limitations. As an answer, the MindMinings research project aims to develop and evaluate a semantic-based profiling system for improvement purpose.

Keywords: Ontology, Big Data, User Profiling, Data Analysis

1. Introduction

The online world has become one of the most important advertising panels. But, in contrast to traditional media, it does not only allow advertisers to put their message out there: The spectrum of the underlying technology enables to target a single user individually. The Web makes, indeed, a big difference: Whereas nobody may know which pages of the morning newspaper are actually read, similar traces are constantly logged and stored when customers surf the internet. The exact pages visited, at what time, for how long, with what device – all those information are automatically available to page owners. Furthermore, cookies, little files left on the hard disk of a page visitor allow web servers to recognize him/her on the next visit and to connect the stored information over time.

There is a fundamental research effort to use these traces for the improvement of services. Personalisation is a central concept to research ambitions in e.g. information retrieval, personal information management and recommendation. Indeed, the problem task that is the focus of the article at hand can be seen as a specialised recommendation task: instead of articles or products, the system facilitates the choice of appropriate advertising contents. However, the features chosen and processed are centred around those relevant for the application and may differ from those very similar tasks.

Driven by these similarities, digital advertising embraces today a mixture of traditional techniques for pattern discovery within a multitude of data. Statistics and machine learning serve to automatically identify customer groups with shared interests and the respective appropriate content. However, so far, no computer system has been developed that perfectly understands a user's background, interests and predicts future intends.

The MindMinings project explores a next step to performance improvement by introducing the ontology-based paradigm to an exemplary online advertising system. By storing all information within a specialised ontology, it will be possible to integrate a multitude of sources.

The ontology model is being developed in close collaboration with a company delivering analysis solutions for digital advertising. It is thus designed to capture the specialised knowledge of domain experts. As

a base for all analysis, it contains information about a user's consulted web resources and, thanks to enhanced analysis, their content and its interpretation. The usage of standardised ontology languages allows furthermore the integration with existing knowledge resources of the Web of Data – a rich set of taxonomies and dictionaries that enhance specialised information with common knowledge (e.g. using WordNet [17], DBPedia [1], or Freebase[3]). The integration of all these information builds up a thorough picture of the user's behaviour and context. The usage of a very fast and lightweight RDF database, coupled OWL 2 and SWRL support [6] allows to use ontological inference to deduce formerly unknown facts that stem from the combination of the above-named knowledge sources.

The article at hand aims to give detailed information about the project context, its keylocks and solutions, as considered and implemented. Section 2 gives an overview of the different goals defined for the project, followed by a survey of related work (Section 3). Section 4 offers thorough description of our system design, the developed ontology and how its functionalities are used for audience segmentation. The article concludes with a summary and an outlook on future work.

2. Task description

The final goal of the MindMinings project is to design a profiling system that uses the navigation traces every user leaves when visiting pages to deduce viable information for digital advertising. As a distinction from similar ambitions, the system will be fully ontology-based, as opposed to traditional statistics- and machine learning-based implementations that dominate today's market.

The CheckSem research team unites researchers on different stages of their careers whose research interests centre around the way semantic web technologies may be applied, adapted and extended for particular applications. To ensure a correct vision on the domain on digital advertising, a collaboration was initiated with one of the French online advertising market actors.

The company offers specialised data analysis to online publishers, editors and advertisers, based on the user data that are available to those parties. From the side of online editors, these information comprise details about the users' navigation on the websites – what pages have been requested, at what time, with what device. With the help of cookies, the visits of a certain machine can be linked over time to constitute a more exhaustive usage profile. In certain cases, when the user deliberately registered to the page, additional information might available that have been explicitly requested on the creation of the user profile – these may comprise basic demographic information, such as age and gender, but also more complex statements as information about topics of interest and purchase intentions.

Even though the nature of the data available to the industrial partner are more varied, the MindMinings system will mainly rely its most stable parts; that is, the users' navigation logs. Information stemming from Customer Relationship Management (CRM) can be used as a measure for validation and evaluation but is, usually, only scarcely available. So the system has to be able to automatically process the newly entering navigation logs and extract the viable information. The logs arrive in a standardised format and can be directly parsed. Contained are basic information about the users' page visits: time stamps, the user agent (a summary about the hardware and software used during the visit), and the exact URL of the displayed contents.

Given this context, the task at hand comes with three key issues:

1. **Resource analysis:** The analysis of the web resources viewed by the user are a crucial step in constructing the user profile. Their content has to be unambiguously analysed in order to be transformed into a machine-interpretable representation.
2. **User profiling:** Based on all available information, a unique profile for each user has to be constructed. The background knowledge about relations between user characteristics has to be extended with knowledge that can be deduced from the user's specific trace.
3. **Big data scale:** Cookies assemble the event traces for the users of each partner site. This leads to a high volume of heterogeneous data that have to be interpreted. Every solution thus has to be efficient and scalable.

Whereas the domains user profiling and resource analysis have been studied for decades, *Big Data* is a domain that emerged during the last years, mainly motivated by the immense amount of data produced by web applications. Indeed, interpreting incoming events along with the dynamic creation of the user profile and its exploitation for content suggestion have to scale to

the number of partner sites and the number of users that surf those pages every day. Thus, the profiling task at hand qualifies in all criteria defined by the IEEE[1] as chracteristics of Big Data, notably:

1. Volume – The amount of data assembled by the company each month reaches about 150 million user events per month for an average partner site, in sum 2,4 billion events each month, leading to a high volume of data to analyse.
2. Variety – Our analysis includes web data in all its formats and orientations, leading to a highly variable spectrum.
3. Velocity – User events arrive for analysis in real-time, meaning a high frequency during main activity hours.
4. Value – As the analysis is performed on real-life data, generated by the activity of users, it bears information about their interests and habits that is commercially exploitable.
5. Veracity – In fast arriving, variably shaped data ambiguities and erroneous entries are not avoidable. The trustworthiness of the source and uncertainty are to be considered in the analysis.

User clicks appear in a high number and also in high frequency. Starting from the numbers indicated above, on an average day, all current clients would produce a number of 80 million data instances scattered over day periods with different degrees of activity. Even though data arrives in form of uniform user events, the web resources referenced therein expose a high level of variety and may include different types of media (articles, blog entries,pictures, videos, etc.) and basically all combinations of topics. The value of the information included, however, is a better understanding of the user behaviour and deduce interests. However, in such a stream, erroneous events will certainly happen. This refers to all click data that are reported, but that are not pertinent for the summary of the user's interests.

Using an ontology to realise the information integration brings manifold advantages. An ontology facilitates the explicit and unambiguous structuring of common knowledge – it allows the clear definition of domain knowledge and thus its reuse throughout applications. The explicit representation makes it possible to analyse the domain knowledge, to draw further conclusions or challenge its implications. In a hierarchical structure, meta-knowledge, domain knowledge and operational knowledge may be separated, allowing a reuse of generic information while keeping the operational information limited to their case of application.

[1] http://www.ieee.org

3. Related Work

Due to the multitude of research domains touched by the MindMinings project, the body of related research is vast. It goes beyond scope of this article to review all literature that proposed solutions for resource and user profiling. The subsequent sections will, however, take a thorough look at works that use ontology-based paradigms for these two tasks. Finally, a short passage sums up efforts to integrate information from different sources based on ontologies.

3.1. Resource Profiling

Algorithms for website qualification comprise techniques for the pre-processing of resources, that extract and normalise the features for each resource; the actual application of classification algorithms and the representation and storage of results.

The widely used technique for the featuring of textual resources is the discovery of keyword terms. Algorithms mostly use statistical considerations to determine the terms within the resource that best summarise its content. Techniques reach from the removal of frequent, non-valuable stop words to enhanced discovery of underlying topic dimensions (Latent Dirichlet Allocation [2]). A simple, but powerful approach that is probably the most used measure for word relevance today is the TF-IDF measure [23]: a mathematical measure that facilitates the discovery of those terms that distinguish a certain document from the rest of the document corpus. Semantic Web resources have been used in this working step as a reference – a large number of works report the success of the inclusion of WordNet [17], DBPedia [1] and specific domain ontologies to the pre-processing of textual resources. [20,21] review techniques for Named Entity Recognition and name structured knowledge sources as a supportive element for the task, besides unstructured sources as corpora. [7] uses a combination of a semantic topic attributed to a term and his syntactic role within the sentence to boost performance for Part-of-Speech tagging. In [18], the inclusion of a Semantic Web reference resource has proven to boost efficiency. The authors compare the traditional statistical LSI approach (Latent Semantic Indexing [8]) for topic discovery with a WordNet-

based alternative. They proved the latter to achieve results as good as the traditional approach, but in considerably lower computation time.

Furthermore, Semantic Web resources have been applied for both, feature reduction and query extension. Both techniques tackle the problem of feature sparseness - given the high number of words in natural language, one text might only feature a small number of the words that are contained in the dictionary of all documents. Besides, the existence of synonyms in natural language might lead to texts that treat the same topic, but show no overlaps in their used vocabulary. In consequence, a regular matching algorithm would not find them to be similar. Anyhow, the background knowledge about semantic relations that is contained in a semantic dictionary as WordNet [17], can be used to either (a) cluster words that are closely related and thus reduce the feature space, or (b) extend each keyword vector by the synonyms of the terms already included and thus facilitate a matching between documents that use disjoint, but synonymic sets of words. However, approaches concentrate around improvements of the traditional keyword based approach. The quality of topic terms benefits from disambiguation or extension by synonymic terms. The semantic relations of the concepts are thus used to ameliorate a traditional paradigm, but then omitted for further exploitation.

3.2. User Profiling

User profiling is a problem task that has been tackled from many directions, the interest in personalisation has been expressed in several domains. The focus will be set on publications that propose ontology-based profiling solutions. The following passages feature a review of meta-models for ontology-based profiling (Section 3.2.1), and resources that facilitate to model complex user interests (Section 3.2.2).

3.2.1. Models for ontology-based profiling

One of the main paradigms that reigns the development of semantic web resources, is to reuse existing repositories as much if possible to avoid duplicate work and redundancy within knowledge resources. A study of literature revealed two main propositions of profile ontology models: the General User Model Ontology (GUMO) [10], and the later publications by Golemati et al. [9].

GUMO: GUMO is an OWL-based representation of user properties and influenced by the already existing specifications provided by UserML, SUMO and the UbisWorld ontology. It adopts UserML's division of the user profile dimensions in three basic classes: auxiliary, predicate and range. Given, one wants to characterise the user's interest in fantasy literature, this might be done using the auxiliary "hasInterest", the predicate "fantasy literature" and the range "low", "medium" or "high". Expressing a statement about knowledge, one might use the auxiliary "hasKnowledge", a predicate that contains the area of expertise that shall be expressed and a range of the statements "poor", "average", "good" and "excellent". GUMO identifies about 1000 groups of auxiliaries, predicates and ranges. Anyhow, one may notice that actually everything can be the predicate for auxiliaries like "hasInterest" or "hasKnowledge", leading to a problem if the ontology design is not realised in a modular fashion [10]. In the case of GUMO, the basic design solution is to limit the ontology to the basic user model dimensions and to leave the modelisation of the general world knowledge open. Thus, one may plug-in an application-specific domain ontology to fill in the slots or integrate a general purpose ontology such as SUMO [22] or UbisWorld [2]. Hence, the missing definition of interest concepts becomes a feature as it enables the simple adaptation to a certain application domain. The basic version as presented in [10] identifies the following set of user model auxiliaries hasKnowledge, hasInterest, hasBelieve, hasPlan, hasProperty, hasGoal, hasPlan, hasRegularity, hasLocation. The listing provided does not aim for completeness and leaves it open to each designer of application-specific adaptations to extend it.

Golemati et al.: Golemati et al. presented their generic user model in 2007. Their identification of re-occurring core concepts in user profiles is based on a thorough literature review. The identified top-level dimensions can be found in Table 1. The central concept of the ontology is the "`Person`" class that contains all the characteristics of the user profile. These information comprise basic facts such as a name or a date of birth; but also instances of the ontology classes as the description of the user's contacts. All classes capturing more complex notions provide slots the respective characteristics of the user's life and a field to represent the time period of actuality for the information. The classes "Interest", "Preference",

[2] www.ubisworld.org

"Ability", "Characteristic" and "Thing" contain three slots: "type", "name" and "score" (or "value" in the case of "Thing"). "Thing" has two sub-classes: "Living Thing" and "Non Living Thing", as adopted from the WorldNet ontology [miller1995]. For the class "Interest" an additional field "interest type" has been included to enable the modelisation of interest classes and their hierarchy.

Remark: It has to be noticed that none of the above examples seems to have an active support community. The links given in the papers are without exception unavailable. However, a rdf- and a owl-version of the approach by Golemati at al. can be found at `http://www.sdbs.uop.gr/?q=node/273`.

UMMO: In a publication from 2005, Yudelson at al. propose a meta-model for user profiling, the User Model Meta-Ontology (UMMO) that aims to specify main streams of the domain and categorise approaches. It serves thus not to structure the terms appearing *within* the various profiling approaches but to categorise the approaches themselves based on the used input data, data structures and algorithms. The keyterms have been extracted from various domain publications and extended and structured based on expert opinions.

3.2.2. Modelling user interests

Both above-named approaches for user modelling leave open slots for the integration of application-dependent concepts for complex characteristics, such as interests, knowledge, etc. Whereas GUMO [10] leaves the decision for a suitable resource up to the ontology designer (but giving a recommendation for SUMO and the UbisWorld ontology), the approach proposed by Golemati et al.[9] includes sub-categories adopted from the WorldNet ontology [17] as a base for extension. Independently from the notions of those basic profile ontologies, several methods provide solutions for the semantic enrichment of identified user interest keywords. The options comprise (a) the relation of the terms with the clearly defined concepts in a reference ontology (mostly used for disambiguation of keywords), (b) the construction of a profile ontology based on extracted parts of the reference ontology and (c) the estimation of relations between the terms without using a reference.

Use of general purpose ontologies: The usage of semantic web resources as a reference has been also applied in user profiling approaches. Often, references to existing knowledge bases are used to disambiguate terms (using WordNet: [25,16,26,4], using ODP: [13]), extend the feature space with their synonyms [25] or relate them to a upper level concept of interest [15].

Extraction of ontology snippets: More recent approaches do not limit themselves to the mere reference of Semantic Web references, but opt to integrate the additional information that they contain to the user profile. The above described approaches use, for example, the synset relation between terms in WordNet to alter the keyword sets for a user/resource. Current methods, however, aim to use all semantic relations surrounding the concept in question. This demands techniques that discover the relevance of the relations and concepts that are connected to a profile term. Calegari&Pasi [5] suggest to use the basic profile terms from a user's collection as seed terms to then collect additional information from a reference ontology. Therefore, the concepts included in the keywords are identified and their neighbouring concepts integrated to the profile, depending on the distance to the original concept and the relation types that exist between them. [15] proposed the Spreading Activation algorithm for the use of semi-automatic ontology extension, [11] shows an application in the context of personal information management. In a later work, Katifori outlines the similarities between the spreading activation used for weight computing on ontologies and current theories about human memory – and suggests its usefulness when designing user-centred information systems [12]. According to him, a model of concept integration and weight adaptation that mimics the capabilities of the human brain will be closest to the personal perception of a user and thus, best able to predict his/her needs. Even though the task-centred profiling incorporated in PIM applications is not exactly equivalent to interest profiling, the approach might be a promising starting point for adaptation.

Generation of knowledge from user's resources: The ontology used for the representation of a user's interest does not need to be connected to a reference resource. Even though the extraction of relations from an existing knowledge repository seems to be a straightforward approach, the properties of the underlying resource collection might suggest the usage of alternative techniques. Given a set of texts that treat a very specific domain of interest, one might not find the necessary content in a general purpose ontology as WordNet. The construction of a specialised domain ontology is a costly process and might not be possible due to financial or timely constraints. In such a case, methods for automatic relation discovery may help to construct

Class Name	Class Description
Person	Basic User Information like name, date of birth, e-mail
Characteristic	General user characteristics, like eye color, height, weight, etc.
Ability	User abilities and disabilities, both mental and physical
Living Conditions	Information relevant to the user's place of residence and house type.
Contact	Other persons, with whom the person is related, including relatives, friends, co-workers.
Preference	User preferences, for example "loves cats", "likes blue color" or "dislikes classical music"
Interest	User hobby or work-related interests. For example, "interested in sports", "interested in cooking"
Activity	User activities, hobby or work related. For example, "collects stamps" or "investigates the 4th Crusade"
Education	User education issues, including for example university diplomas and languages
Profession	The user's profession

Table 1
Upper-level classes of the User Profile Ontology according to [9]

an ontological structure nevertheless. One example for such a automatically generated profile ontology was proposed by the working group around Y. Li and N. Zhong [28,14]. The generated concept relations rely on co-occurrence measures as applied to the user's document collection. From the user's text collection, the most relevant terms are extracted using basic text mining techniques. Using pattern recognition techniques, the discovered terms are grouped and then relationships between the elements estimated using association rule mining. The result is a personalised domain ontology of the user's interest terms. The underlying algorithms used comprise probabilistic classifiers and Hopfield networks [28], fuzzy relational algebra [19] (the latter integrating expert knowledge by letting them attribute semantics to the discovered relations)[5].

4. Our approach

The composition of a comprehensive user profile is a complex task, even if aimed at a limited domain as digital advertising. An intelligent subdivision of tasks is thus indispensable. The main division has already been indicated in the above passages: (a) web resource profiling, (b) user profiling, (c) aggregation and inference. Even though there have been system propositions in literature that treat individual tasks of this list, only few approaches have been found that try an integration of information in form of an ontology. None of these has been specialised for an application domain and, more importantly, tested in an industrial environment.

The collaboration with a company that actually offers profiling services to actors in digital advertising offers a multitude of possibilities for the validation of our approach. The integration of the domain knowledge of the partner is one important point, but a main advantage is surely the possibility to expose the developments to comparative testing and ameliorate the algorithms accordingly.

4.1. General presentation

The core task of the MindMinings project is to design a profiling system that integrates all working steps and data in one structure, the ontology. To ensure the applicability of the approach to a real-world application, all developments have been made in close collaboration with our industrial partner.

The constraints imposed when using the system in a real-world context have important influence on the system design. User events from web platforms arrive in a very high frequency, the system must thus integrate new information in a very fast way – and adapt the respective deductions concerning the advertisement to place for each single user. Even though up-to-date triple stores are designed to come up to such imperatives, the performance requirement do not only include the reaction time of the ontology, but also the semantic analysis process.

To speed up the profiling process that happens following the user activity on the web pages, the semantic resource profiling got decoupled. The company partners up with a limited number of contractual clients that may hold a huge, but still limited, number of websites. On this limited set, the semantic analysis may be done before the actual user activity happens, as soon as a new contract is concluded or a new page enters the ensemble. The idea is to store the already semantically analysed and aggregated information about each web page in the triple store-published ontology and to build up the connections for each user on his/her activity period.

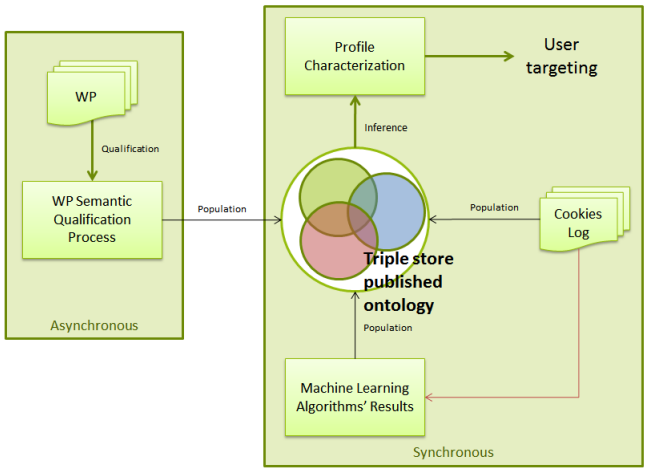

Fig. 1. Overview of the workflow within the MindMinings profiling system

Figure 1 shows an overview of the building blocks of the MindMinings system. The left shows the entering web resources, that are subjected to the asynchronous semantic analysis. The resulting information is used to populate the ontology and will therefore be combined in the synchronous profiling process.

Starting from the log files obtained from a user's browsing behaviour (on the right side of the picture), information are pulled together:

1. facts that can be directly extracted from the usage logs are fed to the ontology (time stamps, user agent etc.);
2. analysis results from the company's machine learning toolkits are brought into ontological form and included;
3. semantic information about the web page links contained in the log are pulled from the storage and linked at run-time.

The combination of all those elements with expert domain knowledge (modelled into the ontology by means of constraints and SWRL rules) about attractive segments and their composition, allows to deduce for each user which of the segments he belongs to and to what degree.

4.2. Ontology: Entity Presentation

Our ontology comprises several classes, all of them being sub-classes of `owl:thing`. The following sections present these entities and their relations, a graphical overview can be found in Figure 2.

4.2.1. Context entities

This category groups the entities that are defined by our industrial partner's 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 domains. Each partner has a different amount information appertaining to him, possibly extended by collaborations with other actors on the market. Hence, information about the partner and his possible coalitions with other players are crucial to determine the facts that are visible or have to be hidden in the analysis process. Hence, the following entities have been included in the ontology:

Partner: The partner is a society that has signed a contract with our industrial partner for the treatment of their data. Each partner is identified by a WID, a "web identification". All domain and web pages that belong to a partner will have this ID attached.

Domain: Referring to the official Domain Name System (DNS), the domain in our 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. In an example: the URL `http://lentrepise.lexpress.fr/index.html` refers to the entry page of the enterprise section of the French journal "L'express". The domain in this case would be "lexpress.fr" (including the top-level domain ".fr" and the second-level domain "lexpress").

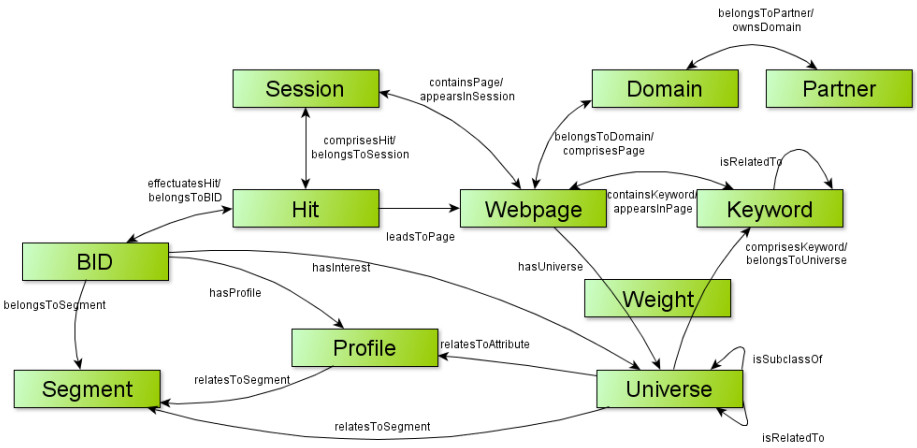

Fig. 2. Graphical overview of the MindMinings ontology

"lentreprise" identifies the sub-domain that is addressed, "index.html" identifies the specific page to display.

Webpage: The class "`Webpage`" envelops all content pages that are found 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 Identifier (URI) that is included in a data property connected to it. The web page is an entity of the company context, but at the same moment a base element for data treatment and analysis. Hence, one may consider it the binding element of those three contextual areas.

4.2.2. Data entities

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

Hit: The hit comprises all information about a single page view of a user. That is, whenever a page is requested from the server, this is logged as one hit. Included in the class are all information related to that entity – the time stamp, the user agent, etc. In the vocabulary of our industrial partner, the information enveloped by the class equals the informational value of the "enriched hit", as we aim to store all information that can be deduced from the basic entities: user location (country, region, town), time spent on the website etc.

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.

4.2.3. Analysis entities

The content analysis process demands the integration of some classes to capture the elements used within. Those mainly comprise the features used to describe the content of a web page, keywords and universes.

Keyword: A keyword is a basic term that describes one concept contained in a web page. The `Keyword` class will be used to capture the keywords found in the web pages and to handle their disambiguation using external knowledge sources. As such they allow the integration of external URIs that link to DBPedia, WordNet or the like. The instances of the `Keyword` class constitute the binding element between a web page individual and the categories ("universes") it belongs to. Given the above definition, we have created the `containsKeyword`-object property which relates a `Webpage` to a `Keyword`. This ojbect property is asymmetric and irreflexive. Its inverse object property is `appearsInPage`. As mentioned above, a keyword belongs to a universe of keywords. Therefore we have created the `belongsToUniverse` object property which is also asymmetric and irreflexive. Its inverse object property is `comprisesKeyword`. We have also defined object properties for modelling se-

mantic relations between keywords, such as antonyms and synonyms.

Universe: The term "universe" stems from our partner's internal vocabulary and 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 keywords that are associated with the respective content domain.

4.2.4. Profile entities

The final goal of all computing efforts is the semantically enhanced profile representation for every website visitor.

Profile: "`Profile`" is the main class linking the attributes making up the user profile. 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 or stem from the analyses' done via statistics and machine learning techniques at the company. 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 or the affinity to certain brands). For the moment, each of those sub-classes is divided in a number of groups signifying the commercially interesting division of the attribute. For example, the age value is currently identified by choosing to link a profile with one 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". The partitions have been chosen based on the formats currently used by the industrial partner.

Segment: One of the key features of the ontology is its capability to automatically determine the attribution of an individual to a certain class. Using the affiliation of a user individual to certain of the above described groups, 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.

4.2.5. Constructional entities

An additional class was added to the ontology for mere internal treatment. The concept `Weight` was added to contain numerical values for each object property that may apply to a certain degree. For a more realistic modelling, it will be convenient to have the possibility to weight the relations between certain concepts. For example, a web page may treat a certain set of topics, but each of them to a different coverage. Hence, we want to model that a certain universe is represented to e.g. 40% in a page and include this notion to the evaluation. As mentioned before, a given user profile will belong to several universes of interest, and a given keyword will appear in several different web pages. We wanted to be able to quantify this degree of belonging. We therefore added the `Weight` concept in our ontology. The instances of this class carry a data type property that contains a numerical value quantifying the weight of the relation. As a direct consequence, the relation between the concept Webpage and the concept Universe, named `hasUniverse`, is specified as being a concatenation of two other relations: `hasWebpageWeight`, relating the concepts of `Webpage` and `Weight` and `weightHasUniverse` that then concludes the relation to the respective Universe concept:

$$hasWebpageWeight \circ weightHasUniverse \; SubPropertyOf \; hasUniverse \tag{1}$$

where `hasWebpageWeight` is forced to connect a `Webpage` with a `Weight`; `weightHasUniverse` to connect a `Weight` with a `Universe`. The like 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).

4.3. Ontology: Segment definitions

There are two ways to define definitions within the current ontology: using class constraints and composing SWRL rules atop classes.

4.3.1. Using OWL restrictions

OWL restrictions are constraints that may be defined on relationships between entities. By their insertion, we define an anonymous class of all individuals fulfilling the constraints, but omit to include this class explicitly in our class hierarchy. This is useful if we want

to distinguish an unnamed set of entities that has no further semantic value for the contents of the ontology and, particularly, if we want the inference engine of our framework to find the affiliations of the individuals automatically.

4.3.2. Quantifier restrictions

A quantifier restriction puts a constraint on a relationship a given individual takes part in. It consists of three parts:

1. a quantifier (either the existential quantifier some or the universal quantifier only
2. the property that is concerned by the restriction
3. a filler that refers to a class

Using those two types of constraints, one can express that at least one kind of this relationship has to exist ("some") or that exclusively this type of relationship may exist.

owl:someValuesFrom: We use the "some"-relation when describing that a certain user shows an interest in certain topics. The corresponding class description would be: "hasVisited some Webpage USports", where "hasVisited" is the relation that connects a user with a web page that appears in his navigation log and "Webpage USports" the class capturing all web pages that treat sports-related topics. By using the "some" relation, it is stated that the user has to have visited at least one page that belongs to those that are related to the universe "Sports", but there is no limit imposed concerning all the other topics that the same web page might be related to.

owl:allValuesFrom: We have chosen to apply the universal restriction "only" in cases when one wants to single out a very specific group of individuals. For example, the following restriction allows identifying only the users that are interested in soccer, but in no other sports: "hasVisited only Webpage USports Soccer".

owl:hasValue: The owl:hasValue restriction is very similar to the owl:someValueFrom constraint, only that the filler is not a class, but a single individual. It thus enables to define an anonymous class of individuals that are connected to one specified individual in the storage. One may note that following the open world assumption, there is no implied constraint concerning the other relations the individuals take part in – they may contribute to other properties, or even the same property but leading to another class of individuals. This kind of constraint becomes useful when targeting specific brackets of the user profile as those are modelled as individuals. For example, to single out the individuals that belong to the age group over 65, it will be enough to define the anonymous class by the restriction "hasAge value Age Over65".

4.3.3. Cardinality restrictions

Cardinality restrictions allow to establish constraints on the number of relations of a certain type the individuals may take part in. They come in three flavours: minimum cardinality, maximum cardinality and cardinality. In compliance with their name, they enable to define a minimal/maximal number of relationships that an individual may have or even set an exact number. As an example, we may specify that a person does only count as being interested in sports, if having seen at least five pages that treat sports-related topics: "hasVisited min 5 Webpage USports".

4.3.4. Using SWRL rules

As presented in section 4.2.5, the weight concept allows us to put a measurement of importance on 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". To actually use the weight notation for reasoning, however, the usage of SWRL rules becomes indispensable. The following example illustrates the definition of a strong relation between a Universe and a Webpage:

$$\begin{aligned} &Universe(?u), Webpage(?wp),\\ &float[>= 0.8, <= 65](?value),\\ &hasUniverseWeight(?w, ?u),\\ &hasWebpageWeight(?wp, ?w),\\ &hasWeight(?w, ?value)\\ \rightarrow\ &hasStrongRelation(?wp, ?u) \end{aligned} \tag{2}$$

Specifying that a universe u and a web page wp that are connected by a concatenation of properties hasWebpageWeight and hasUniverseWeight, connecting them with a weight w with float $value$ that lies between 0.65 and 0.8, it is assumed that wp and u have a strong relation to another.

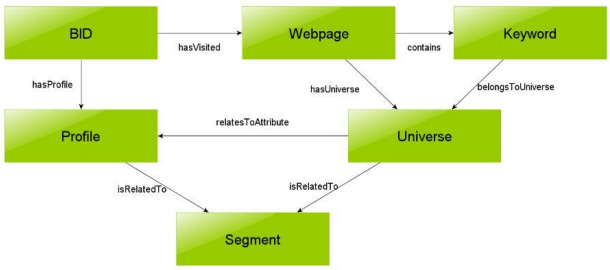

Fig. 3. Minimal version of the ontology used for examples

Fig. 4. The class hierarchy after the creation of the concept "mother"

4.3.5. Practical illustration

The following paragraphs and figures illustrate a prototype application and show some exemplified implementations using the theoretical foundations of above. For the sake of clarity, a minimal version of the user profile ontology was adapted, as visualised in Fig. 3. We mainly omitted the context entities and those data entities that do not directly contribute to the analysis process. However, for the classes included, all specifications obtained during the collaborations have been kept to their maximum. This values particularly for the profile class that contains a rich number of subclasses and their definitions.

The industrial partner aims for marketing-oriented user segmentation. Depending on the commercial partner in question, those segments may vary – a condition that favours the usage of an ontology as flexible underlying data structure. Segments are defined by the commercial division of our industrial partner and are integrated in the ontology to allow a segmentation of users employing inference. The construction of new restrictions and rules will be supported by an intuitive interface that is to be developed in the later stages of the project. Hence, the integration of new segments will be an easy-to-handle process even for personnel not familiar with ontology construction and maintenance. For the moment, individuals belonging each segment have been constructed manually to show the workings of the inference engine. However, please note that the ontology has also been published on a Stardog triple store[6] that works in combination with several web services developed earlier in the project. Thus, it is possible to automatically populate the ontology published on the triple store by handing over a navigation log as initial data structure. The navigation information contained in the logs are parsed automatically and the respective elements (keywords, webpages, sessions, hits, etc.) inserted into the ontology according to the features that were identified by the analysis. Furthermore, it is contemplated to integrate the results stemming from our partner's machine learning internal application as additional source of information. The rest of this section describes how constraints can be used to deduce two segments as they may appear in our industrial partner's definitions. Therein, we choose one basic example that combines attributes using machine learning techniques, and one advanced example that combines those attributes with information stemming from website content analysis.

Basic example For the first example, we chose the example of the segment "mother", designating a person that belongs to a certain age segment, lives in a household with child presence and was identified as being female. All those attributes belong to the basic profile proposition defined by our industrial partner and can today be determined using machine learning techniques. In the ontology, those findings are combined with the evidence stemming from content analysis and in consequence ascertained or corrected.

The segment "mother"is created as a sub-class of the "segment" concept, the class which assembles all concepts that are combinations of base attributes to a higher abstraction level. Figure 4 shows the created segment in Protégé's class hierarchy, . Of course, the software will not automatically know what the segment "mother" depicts in our model and a definition that suits the application context has to be entered manually: A value restriction is used for the class `Gender` (to choose only the individuals that are female), combined with a helper class `Parent` that was defined before as a person of a certain age segment, that lives in a household with child presence. In consequence, the class description changes as shown in Figure 5 – Protégé does not only include the definition that we provided ("mother" being a entity that is a member of "parent" and having a certain gender, in the first line of the information area) but also assigns all definitions

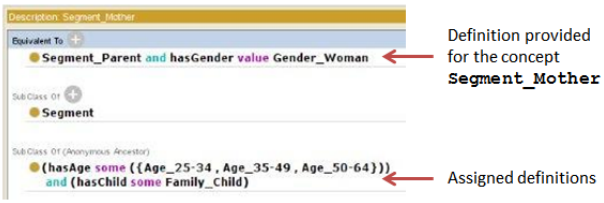

Fig. 5. Changes appearing after having added the definition of the new segment

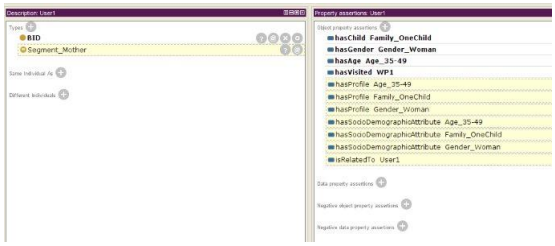

Fig. 6. `Mother` individual with attributes inferred by the reasoner

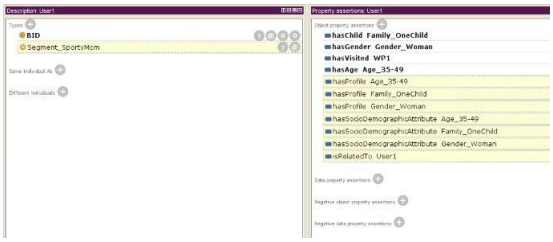

Fig. 7. Changed attributes for the individual "User1" after running the reasoner

that have been entered for the segment parent as implicit super-classes (in the third line of the information area).

For a test, the individual ”user1“ was included to the database and equipped with all the properties of an instance of the class `Mother`. After running the reasoner, the individual is correctly attributed to the respective class – depicted in Figure 6.

Advanced example The same procedure can be realised using attributes not only from the basic profile categories, but also topic information extracted directly from web pages. For the example, again, we construct entities by hand, the user as well as the web page, to exemplify the internal process. However, all those individuals will be discovered automatically from the analysis steps and the ontology populated accordingly. We create a new segment called ”SportyMom”, designating an individual that is a mother (defined by demographic attributes) and has visited some web pages treating topics from the category “Sports“ as follows:

$$Segment\_Mother\ and\ (hasVisited\ min\ 1\ (hasUniverse\ some\ Universe\_Sports)) \quad (3)$$

Furthermore, we create a web page-individual that belongs to the ”Sports” category and define for User1 that she has visited that web page. By the definition of the class, User1 is now not only a mother, but also a “SportyMom“ – a classification that is verified when running the reasoner (e.g. the class attribution for the created indivdual, to be seen in Figure 7).

5. Conclusion and Future Work

In the last eighteen months since the begin of the MindMinings project, the first steps were made to create a thorough, ontology-based user profiling system for the digital advertising domain. Two main achievements were realised in this period:

1. the design of a customised ontology based on the domain knowledge of an industrial partner,
2. the development of a prototype system using current techniques for term extraction, information integration and inference.

The prototype relies on the first version of the ontology that was stored on a triple store. Initial performance test came up to the maximum expected response times, but can certainly be tweaked when using adapted methodology.

The MindMinings ontology is our first contribution to the body of research centring around user profiling. Even though two propositions have been made to standardise profiling processes, none of the projects seems to be still actively used by a community. Furthermore, the entities defined therein do completely fulfil our needs. However, to come up to the semantic web paradigm of resource reuse, one of the next steps will be to connect the concepts used in the MindMinings ontology with their counterparts in the Linked Open Data cloud notably by using `owl:sameAs` predicates to ensure semantic disambiguation of keyword concepts. This will also open access to elements of com-

mon knowledge for our system as it may use the relationships defined within those common knowledge resources for further inference.

The prototype system has been implemented using well-established techniques – as for example the tf-idf measure for the identification of relevant index terms, vector-based distance computation for websites and universes. The next step will be to introduce semantic distance computation for key terms previously disambiguated (by relating them them to the appropriate, clearly defined concept in the Linked Open Data cloud). This customised semantic distance measure will rely on context semantics rather than a mapping of keywords. Furthermore, the basic ontology will be refined and extended to make full use of the weighting concept that allows gradual statements over the affiliations of keywords and web resources to universes, and thus, the degree of a user's interest in a certain topic or product. Adequate frameworks for the propagation of this vague information over the ontology will be evaluated.

The main source of information is composed of textual resources linked to the history of the profiled user. Text analysis has been widely studied through different domains, such as Information Retrieval and Natural Language Processing. Efforts have been made with focus on the extraction of meaning from text and its transformation to machine-interpretable structure. However, the toolkits available often centre around the evaluation of English resources and offer only limited of no support for other languages. As most of the clients of the partner company are French, the adaptation of state-of-the-art approaches to French language will pose one of the mayor keylocks. Furthermore, we will explore extensions of their functionality that serve their customisation for the problem task, focussing on the extraction of information that are significant for market-relevant customer segmentation.

The analysis of natural user behaviour implies another important challenge: human beings seldom act upon logical, crisp principles with clear boundaries and linear reasoning. The profiling system will thus have to handle changing interests, contradictory behaviour or unclear implications. A possible mathematical framework to handle vague and uncertain information could be fuzzy logic, introduced by L.A. Zadeh in 1965 [27]. The inclusion of fuzzy logic to ontological structures has a rather short history, with a first comprehensive work having appeared in 2006 [24]. However, first computing schemes and applications have been presented therein.

All developments will be done with a close eye on the usefulness for the application domain and in search for the best performing alternative concerning the industrial constraints.

References

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM, 2008.

[4] Silvia Calegari and Gabriella Pasi. Definition of user profiles based on the yago ontology. In *IIR, CEUR Workshop Proceedings, CEUR-WS. org*, volume 704, 2011.

[5] Silvia Calegari and Gabriella Pasi. Personal ontologies: Generation of user profiles based on the yago ontology. *Information Processing & Management*, 2012.

[6] LLC Clark & Parsia. Stardog triple store. http://www.stardog.com/.

[7] William M Darling, Michael J Paul, and Fei Song. Unsupervised part-of-speech tagging in noisy and esoteric domains with a syntactic-semantic bayesian hmm. *EACL 2012*, 2012.

[8] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[9] Maria Golemati, Akrivi Katifori, Costas Vassilakis, George Lepouras, and Constantin Halatsis. Creating an ontology for the user profile: Method and applications. In *Proceedings of the First RCIS Conference*, pages 407–412, 2007.

[10] Dominik Heckmann, Tim Schwartz, Boris Brandherm, Michael Schmitz, and Margeritta von Wilamowitz-Moellendorff. Gumo–the general user model ontology. In *User modeling 2005*, pages 428–432. Springer, 2005.

[11] Akrivi Katifori, Costas Vassilakis, and Alan Dix. Using spreading activation through ontologies to support personal information management. *Proc. of Common Sense Knowledge and Goal-Oriented Interfaces*, 2008.

[12] Akrivi Katifori, Costas Vassilakis, and Alan Dix. Ontologies and the brain: Using spreading activation through ontologies to support personal interaction. *Cognitive Systems Research*, 11(1):25–41, 2010.

[13] Georgia Koutrika and Yannis Ioannidis. A unified user profile framework for query disambiguation and personalization. In *Proceedings of workshop on new technologies for personalized information access*, pages 44–53, 2005.

[14] Yuefeng Li and Ning Zhong. Mining ontology for automatically acquiring web user information needs. *Knowledge and Data Engineering, IEEE Transactions on*, 18(4):554–568, 2006.

[15] Fang Liu, Clement Yu, and Weiyi Meng. Personalized web search by mapping user queries to categories. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 558–565. ACM, 2002.

[16] Pasquale Lops, Marco Degemmis, and Giovanni Semeraro. Improving social filtering techniques through wordnet-based user profiles. In *User Modeling 2007*, pages 268–277. Springer, 2007.

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

[18] Pavel Moravec, Michal Kolovrat, and Vaclav Snasel. Lsi vs. wordnet ontology in dimension reduction for information retrieval. *Databases, Texts*, page 18, 2004.

[19] Ph Mylonas, David Vallet, Pablo Castells, Miriam Fernández, and Yannis Avrithis. Personalized information retrieval based on context and ontological knowledge. *The Knowledge Engineering Review*, 23(01):73–100, 2008.

[20] R. Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10, 2009.

[21] R. Navigli. A quick tour of word sense disambiguation, induction and related approaches. *SOFSEM 2012: Theory and Practice of Computer Science*, pages 115–129, 2012.

[22] Ian Niles and Adam Pease. Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, pages 2–9. ACM, 2001.

[23] Gerard Salton and Michael J McGill. Introduction to modern information retrieval. 1986.

[24] Elie Sanchez. *Fuzzy Logic and the Semantic Web*, volume 1. Elsevier Science, 2006.

[25] Giovanni Semeraro, Marco Degemmis, Pasquale Lops, and Pierpaolo Basile. Combining learning and word sense disambiguation for intelligent user profiling. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 2856–2861. Morgan Kaufmann Publishers Inc., 2007.

[26] Giovanni Semeraro, Pasquale Lops, and Marco Degemmis. Wordnet-based user profiles for neighborhood formation in hybrid recommender systems. In *Hybrid Intelligent Systems, 2005. HIS'05. Fifth International Conference on*, pages 6–pp. IEEE, 2005.

[27] Lotfi A. Zadeh. Fuzzy sets, information and cotrol, 8 (3): 338-353, 1965.

[28] Ning Zhong. Representation and construction of ontologies for web intelligence. *International Journal of Foundations of Computer Science*, 13(04):555–570, 2002.