Anett HOPPE, PhD Student
E-mail: Anett.hoppe@ checksem.fr
Ana ROXIN, PhD
Professor Christophe NICOLLE
Checksem Research Group, LE2I
Aile Sciences de L'Ingénieur, Faculté des Sciences Mirande
Université de Bourgogne BP 47870, 21078 Dijon Cedex, France

SEMANTIC USER PROFILING FOR DIGITAL ADVERTISING

Abstract: *With the emergence of real-time distribution of online advertising space ("real-time bidding"), user profiling from web navigation traces becomes crucial. Indeed, it allows online advertisers to target customers without interfering with their activities. Current techniques apply traditional methods as statistics and machine learning, but suffer from their limitations. As an answer, the proposed approach aims to develop and evaluate a semantic-based user profiling system for digital advertising.*

Keywords: *Ontology, Semantic Web, Reasoning, Logical Rules, SWRL, Inference, Big Data, User Profiling, Data Analysis.*

JEL Classification: C81, C55, M37

1. Introduction

The online world has become one of the most important advertising panels, but available techniques for monitoring user's activity on the Web fail in delivering a precise and individual user profile. While detailed information about user navigation (such as the URL of visited pages, time of visit, device used for displaying the website etc.) is captured and available on server-side, it is mainly processed with traditional techniques such as machine learning and statistics. Indeed, these techniques allow to automatically identifying customer groups with shared interests. Most of the traditional machine learning approaches, however, rely on a numerical, vector-shaped input. They thus necessitate the natural language-content of the web resources to be transformed in this format. While there are techniques that achieve that, the transformation often comes along with important losses of semantic information (see Section 3.1.1for details) and direct influence on the quality of results.

The MindMinings project explores a next step to improve performance of profiling for digital advertising. By basing the system on a custom-built domain ontology, we open up the way for the integration of semantically-enriched information. This includes the storage of semantic relationships, and the integration

of existing datasets from the Web of Data. This allows building and delivering a thorough model of the user's behaviour and context. Ontological inference allows deducing formerly unknown facts that stem from the combination of the various information sources and the logical rules contained in the ontology model.

The article at hand aims to give detailed information about the project context, its key locks and proposed solutions. Section 2 gives an overview of the goals defined for the project, followed by a survey of related work (Section 3). Section 4 details the design of our system, section 5 discusses the developed ontology and how its functionalities are used for audience segmentation (Section 6). The article concludes with a summary and an outlook on future work (Section 7).

2. Task Description

Our final goal is to design a profiling system that bases its analyses on users' web navigation logs. The described system is fully ontology-based, as opposed to traditional machine learning-based implementations that dominate today's market.

Furthermore, the system is meant to provide support for experts from digital advertising during the design and implementation of new campaigns. An important part of our work is thus the understanding of the experts' perspective on the concepts and processes of user profiling and how they inter-relate. For this purpose, the knowledge model of the profiling system has been conceived following the guidance of domain experts to model *their* perspective. This allows intuitive integration of the new profiling system to the workflows of a digital advertising company.

We assume only basic information about the user to be known, namely the information that can be recuperated/deduced from the server-side logs that are available to site editors:

- **User Identification:** The user is identified by a key string. If cookies are allowed, this ID will be identical across sessions.
- **User Agent:** contains basic information about the user's technical environment: operating system, used browser, available extensions, device type.
- **Visited Pages:** a list of the URLs that have been visited during this session (or, if available: across sessions).
- **Time Stamps:** every URL is annotated with the time at which it was requested from the server.

However, the overall system design allows for the integration of other information sources. Technological enhancements may lead to additional concepts capturing the results of image or video analysis; furthermore, editors may hold secured data from Customer Relationship Management (CRM), or even introduce new concepts that were unknown during the initial conception of the data structure.

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

1. **Modelling of the Advertising Domain:** Most approaches focus on the automation of user segmentation, assuming computer science experts at both ends of the process. The MindMinings system aims to integrate the perspective of the domain professionals, their terminology and methodology.
2. **Web resource parsing and profiling:** The analysis of the web resources viewed by the user is 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.
3. **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 traces.

The article at hand will focus on two of the problem areas, the conceived model of the advertising and the mechanisms of the user profiling based upon it, respectively.

3. Related Work

The scope of the MindMinings project touches issues from a multitude of research domains and thus, the body of related research is vast. It goes beyond the scope of this article to give a comprehensive review of all literature proposing solutions for web resource and user profiling, and their sub-domains. The subsequent sections will, however, take a thorough look at those research directions that are closest to our focus, namely the use of Semantic Web technology for the profiling of Web resources and users.

3.1.Web page classification

Techniques that serve Web resource qualification comprise algorithms applied in the three stages of processing (a) pre-processing steps, which extract and normalise the features from the resources, (b) the actual application of a qualification method, mostly classification based on the found features, (c) the representation and storage of the results.

The upcoming passages will focus on the first two problem areas. The question for the appropriate data representation is answered by the design of the system surrounding the resource profiling: the results will be fed to the central ontology structure (described in Section 5) for further aggregation and processing. The focus of the following considerations is on the usage of Semantic Web technologies for the purpose.

3.1.1 Pre-Processing

A widely used technique for the featuring of textual resources is the discovery of key terms. The algorithms use statistical considerations to determine the words in the resource best summarise its content. Techniques reach from the

filtering of “stop words” without semantic function to enhanced discovery of underlying topic dimensions (e.g.: LDA (Blei et al. 2003)). The probably most used approach for measuring word relevance is the tf-idf measure(Salton and McGill 1986): a mathematical measure that computes the terms in the text that distinguish a certain document from the rest of the corpus.

These statistical measures use terms that appear directly in the text, and base solely on their frequency of occurrence. This approach is widely used, still today, even though it dates back about thirty years. Reasons lie in its comprehensibility and efficiency. However, they suffer from shortcomings when treating natural human language. The assumption underlying techniques like the tf-idf measure is that texts treating similar topics will use the same vocabulary. Human language, however, is very rich. Using synonymous terms, two texts can treat the same topic – without using vocabulary that overlaps much. The other way around, in the case of homonyms, the vocabulary can show strong overlaps without indicating closely related semantic topics: See the sentences “We grilled fresh bass on the barbecue.” and “He was the one playing the bass in the band at the barbecue.” Both of them contain the word “bass”, but using it to refer to different real-world concepts (the species of fish in the first case, the music instrument in the second).

Semantic Web resources have been used as references, trying to bridge the gap between real-world concept and string chain. A number of works report success by including WordNet(Miller 1995), or DBPedia (Auer et al. 2007). (Navigli 2012)names structured knowledge resources as a viable reference for Named Entity Recognition, (Darling et al. 2012) integrates them to boost performance in Part-Of-Speech tagging. Other works experiment with the reduction or extension of feature sets using external knowledge resources(e.g. (Gong et al. 2005)).

3.1.2 Resource Classification

The above mentioned techniques for pre-processing output an internal representation of the Web resources – usually a set of key words/key concepts with a frequency or importance score attached. To classify those quantitative data to a set of known semantic categories, any arbitrary algorithm for classification is applicable. However, due to the attributes of natural language, the resulting feature sets are often sparse and high-dimensional – the applied classification technique should be adapted to this fact. Additionally, the results of semantic analysis can be quite noisy, another point that should be taken into account when choosing an appropriate algorithm.

For its easiness and efficiency, the Naive Bayes Classifier has been one of the popular algorithms (Yang et al. 2002), as well as the K-Nearest-Neighbour algorithm (Yang et al. 2002) and decision trees(Hasan and Rahman 2003). Combined with rigorous feature selection, Support Vector Machines got applied

(Sun et al. 2002). All of them use the vector space model (Salton et al. 1975) to compute the similarity between two documents.

Web documents are interconnected by hyper-links which give further indication about content relationships (Sutton and McCallum 2006).Relaxation labelling is one relational learning algorithm that has been widely applied to web resource classification. Web pages are first labelled with category memberships based on their content, then relabelled based on the category assignments of their neighbours(Chakrabarti 2002).

As pages are likely to not only discuss one distinct topic, several researchers experimented with soft computing: (Zhou and Huang 2009), for instance, examines e-Commerce web sites using a fuzzy framework. (Espejo et al. 2010) studies the usage of genetic programming for text classification.

The usage of ontologies directly for the classification of documents has been explore in papers like (Song et al. 2006). A recent work (Werner et al. 2014) discusses multi-label classification based on ontological inference. The authors describe how the ontology-based approach facilitates complex classification tasks while offering a clear justification of the classification decision, using the vocabulary of the domain experts.

Implications for the MindMinings project: While there is a rich body of research on how to classify Web resources in pre-defined topic categories, Semantic Technology has not yet been fully integrated in the process. Many approaches explore semantic enhancements of keyword sets – however, for the classification they often fall back to well-known classification techniques. Doing so, they drop additional semantic information, as for example information about relationships among the key-terms as traditional algorithms cannot handle that complex input data.

The MindMinings system will use similar techniques for the initial discovery of key concepts in the textual resources; but lay an important focus on the semantic enhancement of the representations, via disambiguation and relation discovery. The resulting, more complex resource profile is hard to align with the input formats of machine learning techniques. Similar to (Werner et al. 2014), the resource profile will be integrated in the central knowledge base of the system, the MindMinings ontology. The classification of contents will then be realised using the inference engine of the ontology, using the entirety of semantic knowledge about the resource.

3.2.User profiling based on Semantic Web technologies

3.2.1 Standardised Ontology Models for User Profiling

One of the paradigms that govern the development of Semantic Web resources is to reuse existing repositories to avoid duplicate work and redundancy within knowledge resources. However, despite the importance of user modelling for information systems, the number of propositions for a standard model for user profiling is rather disappointing. Anyhow, the found three resources (Heckmann et

al. 2005, Golemati et al. 2007, Ghosh and Dekhil 2009) link to their repositories, but to the best of our knowledge, none of them is maintained up to today.

There were several reasons to abstain from the re-usage of one of these resources, but the missing community support is certainly the most important one. The description of GUMO(Heckmann et al. 2005) is detailed and allows conclusions about its main structure, but it is impossible to recuperate all of the 1000 classes that it supposedly contains. The support of graded statements as we envision to include them in the MindMinings ontology was either weak (Heckmann et al. 2005), or not implemented at all (Golemati et al. 2007).

Furthermore we found, that domain experts from digital advertising hold a profoundly different perspective on user profiling than it was proposed by these generic user modelling propositions. The included concepts did not resonate with our contact persons, a discordance that ranged from simple unintelligibility to missing concepts. In consequence, we decided to start our knowledge model from scratch. It allowed us to stay as close to the understanding of the advertising experts as possible, while leaving us the option to map our concepts to those in other profile ontologies (and general purpose ontologies) in a post-processing step.

3.2.2 Automatic construction of ontology-shaped profiles

Overlay models over domain ontologies: Several works use an explicit model of their problem domain to construct the user profiles. All users share the domain ontology as basic profile structure. The profiles differ by weights assigned to each of the concepts, constituting a so-called overlay model. In the case of large domain ontologies and/or lots of users, this approach necessitates huge amounts of storage. (Middleton et al. 2004) proposes a recommender system for scientific publications based on this approach, (Cena et al. 2011)addresses the modelling of gastronomic preferences and lay a particular focus on update mechanisms within overlay profiles.

Extraction of ontology snippets: More recent approaches do not limit themselves to referencing Semantic Web references, but integrate the additional information that they contain to the user profile. These methods aim to use semantic relations surrounding relevant concepts. This demands techniques that determine the importance and relevance of the relations and concepts surrounding a profile term. (Calegari and Pasi 2012) suggests 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 expressed by the found keywords are identified and they and their neighbouring concepts and relations are integrated to the profile. (Liu et al. 2002) proposed the Spreading Activation algorithm for the use of semi-automatic ontology extension, (Katifori et al. 2008) shows an application in the context of personal information management.

Implications for the MindMinings project: The above sections discussed works that are closely related to the goals of our project. In the spirit of re-usage, we tried to identify possible existing knowledge models for user profiling.

However, the models that were found in literature are neither maintained to the present day, nor do they include all the expressiveness that we aim to equip our system with.

In what concerns the automatic profiling of users based on ontologies, the MindMinings system shows resemblance with the publications naming overlay models – the user's properties will be represented with the help of weights attached to the contained relationships. In contrast to other propositions, the MindMinings system integrates all surrounding steps for Web resource qualification and uses external knowledge resources to extend the information based on semantic relationships. Furthermore, the system is specialised to the specific profiling task for digital advertising, opposed to former, more generic infrastructures.

4. Components of the Profiling System

The composition of a comprehensive user profile is a complex task, even if performed with a limited scope 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.

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 high frequency, the system must integrate new information rapidly– 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 does not only include the reaction time of the ontology, but also the semantic analysis process.

To speed up the profiling process during the user's surfing session, the resource profiling was separated from the actual user profiling. Our reference enterprise works with a limited number of contractual clients that, for their part, administer a limited number of websites. The semantic analysis can be done before the actual user activity happens, as soon as a new contract is signed or by monitoring new arriving pages. The semantically analysed and aggregated information about each web page is stored within the central ontology. During the activity period of the user, his/her profile is created by adapting the connection structure in the ontology.

Figure 1 shows an overview of the building blocks of the MindMinings system. The left shows the entering web resources, which are subjected to semantic analysis. The resulting information is used to populate the ontology and will be combined during the profiling process on run-time.

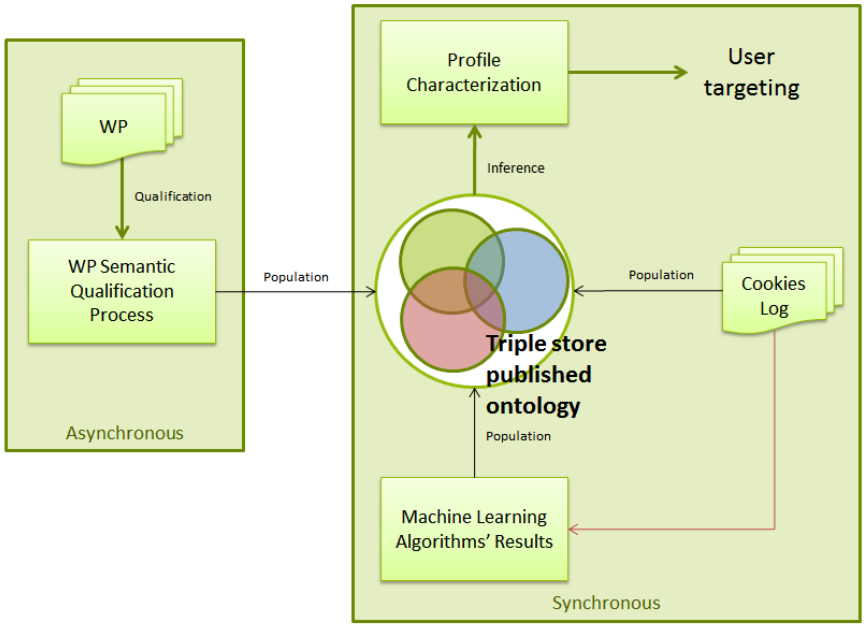
Starting from the log files obtained from a user's browsing behaviour (on the right side of the picture), information are pulled together:

1. Facts that can be directly extracted from the usage logs are fed to the ontology (time stamps, user agent etc.);

2. Semantic information about the web page links contained in the log are pulled from the storage and linked at run-time.

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

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



5. The MindMinings Ontology

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.

5.1. Context entities

This category groups the entities that are defined by the industrial ecosystem surrounding the profiling system in digital advertising.

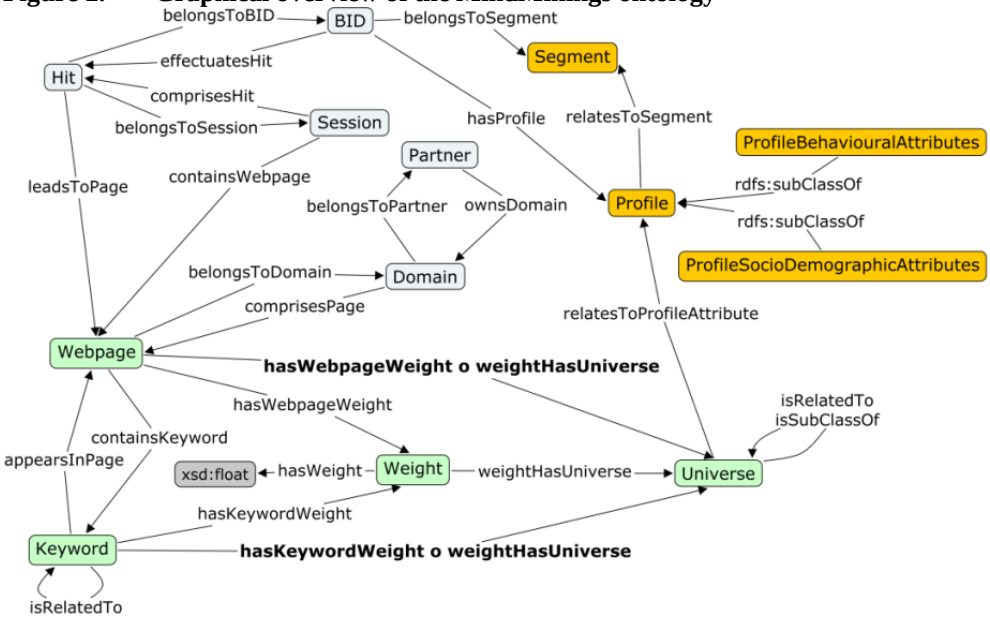
The company concludes a contract with online publishers to provide enhanced analysis of their usage logs. The available information is not only determined by the new client himself .Each partner has different amounts of information appertaining to him, possibly extended by collaborations with other actors on the market. Hence, information about the partner and his possible coalitions with other players is crucial to determine the facts that are visible or

have to be hidden in the analysis process. Thus, the following entities have been included in the ontology:

Partner: A partner is a company that has signed a contract for the treatment of their data. Each partner is identified by a WID, short for "web identification". All domains 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. As 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"). "lentreprise" identifies the sub-domain that is addressed;"index.html" identifies the specific page to display.

Figure 2. Graphical overview of the MindMinings ontology



Webpage: The class "Webpage" envelops all content pages that are found for a certain (sub-) domain. Every parsed web page is an individual, belonging to the class Webpage, and relating with other entities that semantically 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.

5.2.Data processing-related entities

The entities in this section provide additional contextual information to the web pages extracted from the navigation logs. 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 semantic evaluation (as for the entities in Section 5.3).

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.

5.3.Analysis-related entities

During the content analysis process, webpages are qualified with respect to their content. The concepts necessary to capture this information are described below.

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. The contains Keyword-object property relates Webpages to Keywords. Every Keyword belongs to one or more Universes (see below) of Keywords that specify the topic category it belongs to. Furthermore, relations are defined, that allow to specify inter-keyword relations such as antonymy and synonymy.

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.

5.4.User profile-related entities

The final goal of all computing efforts is the semantically enhanced profile representation for every website visitor. Its lower-level features are captures in the super-class "Profile", higher-level compositions as sub-classes of "Segment".

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. 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", or "Age Child". The partitions were chosen based on the formats currently used by the interviewed domain experts.

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 class "Segment" captures these more complex profile entities based on profile features ("a female person living in a household with children" belongs to the segment "mother"), content features ("a person reading 90% of time on sports-related pages 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.

5.5.Constructional entities

An additional class was added to the ontology for 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 to a different degree. 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. Using the Weight concept we will be able to quantify this degree of belonging. The instances of this class carry a data type property that contains a numerical value quantifying the weight of the relation.

A graded relation is specified using a concatenation of two other relations. For instance, the relation among the concepts Webpage and Universe, named hasUniverse will is as follows: hasWebpageWeight, relating the concepts Webpage and Weight and weightHasUniverse that concludes the relation by linking the Weight with the respective Universe concept:

$$hasWebpageWeight \circ weightHasUniverse \subseteq hasUniverse,$$

where hasWebpageWeight is forced to connect a Webpage with a Weight; weightHasUniverse to connect a Weight with a Universe.

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).

6. Using the Ontology for User Profiling

The ontology language OWL does not only allow defining the conceptual structure of the domain. Its expressiveness also allows clarifying which kind of concepts/individuals can be connected by a certain semantic relationship and even how many entities can take part in it. Therefore, these functionalities offer a comfortable and intuitive way to define profile attributes and higher-level segment definitions, using the natural vocabulary defined in the ontology.

The upcoming paragraphs will first describe the nature and the expressiveness of the used features of the OWL language; then show in an example how these functionalities are used to define actual user segments.

6.1.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.

6.1.1 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 (e.g. existential quantifier some, 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 sports-related web pages. 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 the other topics that the same web page is 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 the users that are

interested in soccer, but in no other sports: “hasVisited only Webpage USports Soccer”.

owl:hasValue: The owl:hasValue-restriction constraints the relation to be connected to one single individual. It thus enables to define an anonymous class of individuals that are connected to other specified individuals. This constraint is useful when targeting specific brackets of the user profile as those are modelled as individuals. For example, to single out individuals that belong to the age group over 65, one needs to define the anonymous class “hasAge value Age Over65”.

6.1.2 Cardinality restrictions

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

6.2.Using SWRL rules

As presented in Section 5.5, the weight concept allows us to put a degree of belonging 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”.The following example illustrates the definition of a strong relation between a Universe and a Webpage:

$$Universe(?u), Webpage\ (?wp), float[\geq 0.8, \leq 0.65](?value),$$
$$hasUniverseWeight(?w, ?u), hasWebpageWeight(?wp, ?w),$$
$$hasWeight(?w, ?value) \rightarrow hasStrongRelation(?wp, ?u)$$

The rule specifies that the if universe U and the webpage WP are connected by a concatenation of the properties hasWebpageWeight and hasUniverseWeight, via an individual W of type Weight with an attached float value between 0.65 and 0.8, U has a strong relation with WP. The specification of such fine-grained concepts needs SWRL rules in addition to the functionalities of the OWL language.

6.3.Practical illustration

Advertising campaigns are constantly changing – surely, the products marketed, but also the target audience which is to be reached with a certain ad. The definition of new target segments has thus to be simple enough to be tackled by personnel without background in ontology engineering.

The first step to achieve the necessary transparency is the “lifelike” model that reproduces the process of advertising profiling in the central ontology. In a further step, graphical user interfaces have been proposed that allow an intuitive manipulation of the concepts in the ontology for the construction of flexible customer segments (Figure 3). The left side of the screen shows the concepts in the ontology. Dragging them from the list and dropping them into the rule editor, the user can construct the new segment – using constraints and restrictions. The rules are directly pushed to the ontology, and are integrated to the inference process.

The following paragraphs illustrate the functioning of the ontology inference based on some exemplified segment definitions. For demonstration, the individuals and segment definitions have been manually entered. However, the ontology has been published on a Stardog triple store[1]and is automatically populated from the user navigation logs that arrive in form of JSON-files. The navigation information contained in the logs is parsed automatically and the respective individuals (users, webpages, keywords etc.) inserted into the ontology.

Figure 3. Graphical user interface for the creation of SWRL rules

Basic example: For the first example, we chose the example of the segment "Mother", designating a person that is female, belongs to a certain age segment, and lives in a household with child presence. The deduction of these facts can be based on various sources: the most reliable is potentially available CRM information that can be integrated to the profiling process. Account data, for example, could supply age and gender of the user. On the other hand, if these data are not available, the analysis will solely rely on content information.

[1]http://www.stardog.com

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 semantic meaning of the term has still to be defined: 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 an 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 that have been entered for the segment parent as implicit super-classes (in the third line of the information area).

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

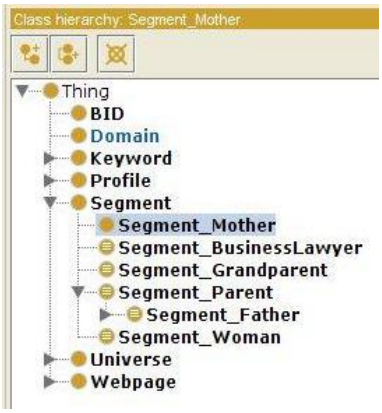
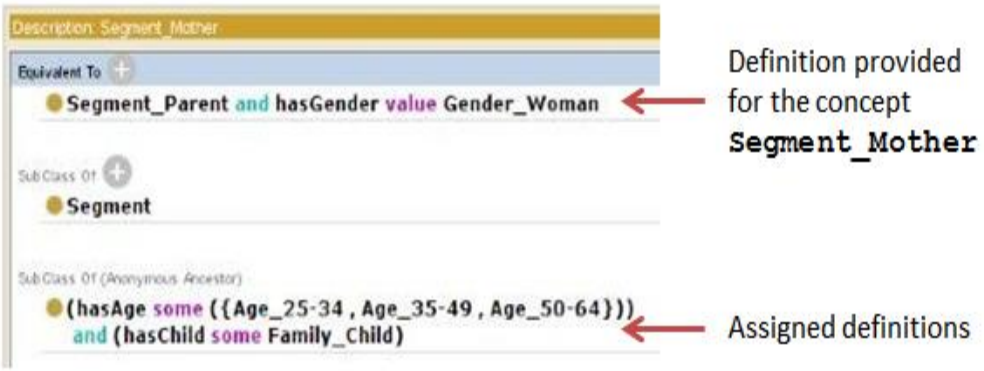


Figure 5. Automatic assignment of implicit segment definitions



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. The black-on-white features are the statements originally made about the individual, while the yellow-shadowed are those that were deduced by the reasoning process.

Figure 6. Mother individual with attributes inferred by the reasoner

Advanced example: The same procedure can be applied using attributes not only from the basic profile categories, but also topic information extracted from web pages. 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:

$$Segment_{Mother} \cap (hasVisited \min 2\ (hasUniverse\ some\ Universe_{Sports})$$

Figure 7. Inferred attributes for the individual "User1" after running the reasoner

Furthermore, we create a web page-individual that belongs to the "Universe_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 individual, to be seen in Figure 7).

7. Conclusion and Future Work

Since the beginning of the MindMinings project, a first prototype was developed for a thorough, ontology-based user profiling system for the digital advertising domain. Two main achievements were realised in this period:

1. design of a customised ontology based on the domain knowledge of experts,
2. development of a prototype system using state-of-the-art techniques

The MindMinings ontology is our first contribution to the body of research around user profiling. Even though propositions have been made to standardise profiling processes, none of themseems to be still actively maintained. Furthermore, the defined concepts do not fulfil our needs in expressiveness. However, one of the next steps will be the necessary connection of the concepts in the MindMinings ontology to their counterparts in the Linked Open Data cloud notably by using owl:sameAspredicates to ensure semantic disambiguation of keyword concepts. This will also open access to elements of common knowledge for our system as it may use the relationships defined within those common knowledge resources for further inference. The ontology has been modelled in a modular way, by grouping concepts that are related to the same semantic context unit (e.g. by assembling all concepts related to the lower-level profile below one super-concept "Profile", by summing up higher-level profile characteristics as sub-classes of "Segment"). This ensures the system's portability to other contexts, where the profiling functionality is needed, but not the adaptations to digital advertising.

The prototype has been implemented using well-established techniques – e.g. the tf-idf measure for the discovery of relevant key terms and vector-based distance computation. The next step will be to introduce semantic distance computation for key terms previously disambiguated (by relating 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 terms.

Furthermore, the basic ontology will be extended to make full use of the Weight-concept to allow gradual statements about the included concepts, 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 (as, for instance (Katifori et al. 2008)).

One mayor issue for future developments is the limited availability of text analysis frameworks for French language: Many toolkits propose reliable

evaluation of English resources; however, the support for other languages is rather limited. As most of the clients of the partner company are French, the adaptation of state-of-the-art approaches to French will pose one of the mayor key-locks.

REFERENCES

[1] **Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives (2007),** ***Dbpedia: A Nucleus for a Web of Open Data.*** Pages 722–735 The semantic web. *Springer*;

[2] **Blei, D. M., A. Y. Ng, and M. I. Jordan. (2003),** ***Latent Dirichlet Allocation.****The Journal of Machine Learning research* 3:993–1022;

[3] **Calegari, S. and G. Pasi (2012),** ***Personal Ontologies: Generation of User Profiles Based on the YAGO Ontology.*** *Information Processing & Management;*

[4] **Cena, F., S. Likavec and F. Osborne (2011),** ***Propagating User Interests in Ontology-based User Model.*** Pages 299–311 AI* IA 2011: *Artificial Intelligence Around Man and Beyond. Springer;*

[5] **Chakrabarti, S. (2002),** ***Mining the Web: Discovering Knowledge from Hypertext Data***. Morgan Kaufmann;

[6] **Darling, W. M., M. J. Paul and F. Song (2012),*****Unsupervised Part-of-speech Tagging in Noisy and Esoteric Domains with a Syntactic-semantic Bayesian hmm***. EACL 2012;

[7] **Espejo, P. G., S. Ventura and F. Herrera (2010),*****A survey on the Application of Genetic Programming to Classification***. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 40:121–144;

[8] **Ghosh, R. and M. Dekhil (2009),*****Discovering User Profiles.*** Pages 1233–1234; Proceedings of the 18th international conference on World wide web. ACM;

[9] **Golemati, M., A. Katifori, C. Vassilakis, G. Lepouras and C. Halatsis (2007),*****Creating an Ontology for the User Profile: Method and Applications*****.** Pages 407–412; Proceedings of the First RCIS Conference;

[10] **Gong, Z., C. W. Cheang and U. L. Hou (2005),*****Web Query Expansion by Wordnet.*** Pages 166–175; Database and Expert Systems Applications. *Springer;*

[11] **Hasan, M. M. and C. M. Rahman (2003),*****Text Categorization Using Association Rule Based Decision Tree*****.** Artificial Intelligence 5:2;

[12] **Heckmann, D., T. Schwartz, B. Brandherm, M. Schmitz and M. von Wilamowitz-Moellendorff (2005),*****Gumo–the General User Model Ontology.***Pages 428–432; User modeling 2005. *Springer*;

[13] **Katifori, A., C. Vassilakis and A. Dix (2008),*****Using Spreading Activation through Ontologies to Support Personal Information Management.*** *Proc. of Common Sense Knowledge and Goal-Oriented Interfaces*;

[14] **Liu, F., C. Yu and W. Meng (2002),*Personalized Web Search by Mapping User Queries to Categories.*** Pages 558–565; *Proceedings of the eleventh international conference on Information and knowledge management.*ACM;
[15] **Middleton, S. E., N. R. Shadbolt and D. C. De Roure. (2004),*Ontological User Profiling in Recommender Systems.*** ACM Transactions on Information Systems (TOIS) 22:54–88;
[16] **Miller, G. A. (1995),*WordNet: A Lexical Database for English.*** Communications of the ACM 38:39–41;
[17] **Navigli, R. (2012),*A Quick Tour of Word Sense Disambiguation, Induction and Related Approaches*.** SOFSEM 2012: Theory and Practice of CS:115–129;
[18] **Salton, G. and M. J. McGill (1986),*Introduction to Modern Information Retrieval.***
[19] **Salton, G., A. Wong and C. S. Yang (1975),*A Vector Space Model for Automatic Indexing.*** Communications of the ACM 18:613–620;
[20] **Song, M.-H., S.-Y. Lim, S.-B. Park, D.-J. Kang and S.-J. Lee (2006),*Ontology-Based Automatic Classification of Web Pages*.** Pages 483–493 *in* A. Abraham, B. de Baets, M. Köppen and B. Nickolay, editors. Applied Soft Computing Technologies: The Challenge of Complexity. *Springer Berlin Heidelberg*;
[21] **Sun, A., E. P. Lim and W. K. Ng (2002),*Web Classification Using Support Vector Machine.***Pages 96–99; Proceedings of the 4th international workshop on Web information and data management. ACM.;
[22] **Sutton, C. and A. McCallum (2006),** ***An Introduction to Conditional Random Fields for Relational Learning.*** Introduction to statistical relational learning. *MIT Press;*
[23] **Werner, D., N. Silva, C. Cruz and A. Bertaux (2014),*Using DL-reasoner for Hierarchical Multilabel Classification Applied to Economical e-news.*** Pages 313–320; Science and Information Conference (SAI), 2014.IEEE;
[24] **Yang, Y., S. Slattery and R. Ghani (2002),** ***A Study of Approaches to Hypertext Categorization.*** *Journal of Intelligent Information Systems* 18:219–241;
[25] **Zhou, D. and W. W. Huang (2009),** ***Using a Fuzzy Classification Approach to Assess e-commerce Web sites: An empirical investigation.***ACM Transactions on Internet Technology (TOIT) 9:12.