Dynamic, Behavior-Based User Profiling Using Semantic Web Technologies in a Big Data Context

Anett Hoppe, Ana Roxin, and Christophe Nicolle

CheckSem Group,
Laboratoire Electronique, Informatique et Image,
Université de Bourgogne
Dijon, France
http://www.checksem.fr

Abstract. The success of shaping the e-society is crucially dependent on how well technology adapts to the needs of each single user. A thorough understanding of one's personality, interests, and social connections facilitate the integration of ICT solutions into one's everyday life. The MindMinings project aims to build an advanced user profile, based on the automatic processing of a user's navigation traces on the Web. Given the various needs underpinned by our goal (e.g. integration of heterogeneous sources and automatic content extraction), we have selected Semantic Web technologies for their capacity to deliver machine-processable information. Indeed, we have to deal with web-based information known to be highly heterogeneous. Using descriptive languages such as OWL for managing the information contained in Web documents, we allow an automatic analysis, processing and exploitation of the related knowledge. Moreover, we use semantic technology in addition to machine learning techniques, in order to build a very expressive user profile model, including not only isolated "drops" of information, but inter-connected and machine-interpretable information. All developed methods are applied to a concrete industrial need: the analysis of user navigation on the Web to deduct patterns for content recommendation.

1 Introduction

The emergence of the World Wide Web has shaped our lives like hardly any invention of the last century. The way humans perceive and process information has been altered and is still evolving based on new technologies each day. The last decade has been marked by a evolution of the paradigm underlying web development. There is a new vision of Web facilities as ubiquitous, human-centered technologies, designed to assist individual information needs in every life situation. For the first time, in online applications, we have the opportunity to target an individual user, not only to a rough group of consumers.

This vision has to be supported by a thorough understanding of the forces that drive a user, by the means of data that are perceived from him and not

Y.T. Demey and H. Panetto (Eds.): OTM 2013 Workshops, LNCS 8186, pp. 363–372, 2013.

intervening with his normal course of action – and thereby treating his privacy with responsibility. The research work presented in this paper describes a novel approach for user profiling based on his behavior observed by a server-side application.

This complex application associates numerous research domains – e.g. natural language processing and semantic annotation for the analysis of the related web sites; knowledge bases for the integration of the multi-shaped information sources along with their organization and exploitation. Following the scope of this workshop, we limit the focus here to the envisioned insertion points for Linked Open Data to our system.

2 Project Context

The MindMinings project is set up in the context of a collaboration between the Ezakus Labs HQ (based in Bordeaux, France) and the CheckSem research group from the University of Burgundy, Dijon, France. The following paragraphs aim to provide an overall vision on the project, its context and the arising challenges. The industrial partner is a two years-old enterprise that settles in the online advertising sector. Specialising the offer to the needs of the digital advertisement ecosystem, it provides analysis solutions for enhanced customer targeting.

Our profiling approach will be integrated to the company current structures. For such, it is essential to capture the implicit knowledge of the professionals within the enterprise in an explicit, ontological representation, and to offer the possibility to integrate already existing, well-established procedures and their results. Apart from that, the main focus will be the realisation of a new module of semantic analysis that shall complement the currently used techniques.

The company offers its services to all actors within the digital advertisement ecosystem. Among its clients, it counts some of the mayor French online publishers. In consequence, the amount of data that has to be processed on a continuous basis is immense and growing. Each user event registered on a partner site is submitted in real-time, summing up to about 2.4 billion user events per month. These events, collected in navigation logs, contain all facts that are known from a user as they can be retrieved by his activity and that are contained in cookies.

At the current state, Ezakus uses enhanced machine learning and data analysis techniques to predict user interests relevant to the targeting purpose. Our ontology-based approach aims to complement this approach, (a) by integrating existing results and methods, (b) by extending them with the help of semantic-based analysis, (c) ontological-inference for the deduction of the user's final profile.

All these analysis' have to be realized within the constraints of the real-world application. That is, be able to scale up to immense and variously-shaped amounts of data and deduct information in reasonable time. Given the orientation to give suggestions of advertisements within the user session, this demands real time-like performance. The extraction of knowledge from textual resources is a complex task, therefore, a semantic analysis on huge amounts of data might exceed the time available. Hence, we split the necessary tasks into online and

offline components. The websites that have to run through semantic analysis are a huge set of data – but limited to the client domains of Ezakus. Thus, we will be able to run the analysis those textual data in an offline process and store the results in the ontology – leaving only the combination of the contained information and the deduction of the user profile to the online process. Figure 2 shows a schematic overview of the information sources that are used to populate the ontology and their splitting into online and offline processes.

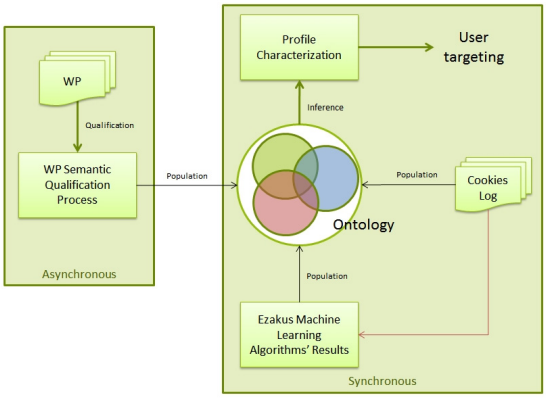

Fig. 1. Overview of the information that will be used to populate the ontology

It is notable that in the current developments, we work in a mostly French environment. Even though the client profiles to deduct are very similar for Western countries, the main language of the sources that we analyze leads to additional issues. The most available frameworks for the treatment of natural language are focused on the analysis of English resources and offer only limited functionalities for other. The toolkit OpenCalais[1] for example offers the discovery of more than 120 named entities and events from English text, but only a set limited to 15 named entities for French language. The adaptation of techniques for their usage on French resources will thus constitute one important branch of the development work.

3 Related Work

As shown in the section describing the context of the project (Section 2), various working steps have to be accomplished. Naturally, research works directly concerned with user profiling will form the starting point of the analysis. However, given the dynamic environment, various other steps will have to be executed automatically and without supervision. We focus here on the tasks concerning information integration and external data sources querying to emphasize the role

[1] `www.opencalais.com`

of linked open data for the integration of a multitude of heterogeneous information sources.Current systems mainly differ in their combinations of techniques for the different stages of the profiling process, (a) information gathering, (b) data representation, (c) exploitation, and (d) updating.

3.1 Information Gathering

The most popular classification of information gathering methods was introduced by Gauch et al. [3]. They propose a categorization based on the degree of active involvement demanded from the user. On the one end of the scale, explicit information gathering refers to the direct request of facts from the user. Therefore, surveys or questionnaires may be used and directly related to a unique user ID, one may also ask for feedback about the previously provided resources (indicating if a certain resource was "interesting" or not). On the other extreme, implicit methods do not directly involve the user, but gather information based on observation of behavior. Instead of asking an explicit opinion about a resource, one may accumulate all data about the resources displayed (content, time stamp of the request etc.), but also access and analyze stored information such as emails or calender events [5].

The usage log analysis, such as envisioned in the MindMinings project, falls into the category of implicit information gathering. As we will only access server-side data, the privacy concerns connected to the analysis of personal data are omitted. Similar approaches were proposed in personalized information retrieval [5], adapting search results by an analysis of the click-through behavior and queries [15,10].

3.2 Information Representation

Representations for user profiles can be classified in three main groups according to [3]: keyword-based, semantic network-based and concept-based techniques. Each of the techniques employs a different basic unit for knowledge representation that affect the amount of information that is exploitable later on [1]. Keyword-based approaches use a set of terms (often weighted) to summarize the user interest. The main advantage of these techniques is their simplicity; however, they tend to reach the limits of their expressiveness when coming across certain particularities of natural language, such as homonyms and synonyms.

Semantic networks were designed to tackle this issue: Concepts are represented by nodes, whereas the arcs in between are established based on the analysis of the user information. Entering information enriches the profile with new keywords that get interconnected with existing nodes [4]. As the developed algorithms did not come up to the expectations [8,12], concept-based approaches are on the rise. By relating the discovered terms with predefined semantic resources, one is able to make their sense machine-processable.

In recent works, several approaches have been presented, using different semantic resources for the identification of concepts. Several approaches use WordNet to adapt their features spaces [2]. Other use light-weight ontologies such as the Yahoo Web Directory [6] or the Open Directory Project [14]. Methods vary

in their usage of the ontological reference – from using the high-level concepts to map the resources to a clearly defined category [7,13] to the extraction of ontology parts and their integration into the profile ontology [1]. In contrast to former approaches, the latter uses YAGO as reference, criticizing former approaches for their adaptation of light-weight frameworks such as WordNet and ODP [9].

4 Project Tasks

As noted in the Introduction section, the overall vision of the project relies on the following elements: a comprehensive model of the user, including estimations of his mayor interests and deductions on his customer behavior. The goal is ambitious as the only input information the analysis relies on are the navigation logs captured while a user browses a publisher's website.

4.1 Web Page Analysis

The server-side information, the log file, will be parsed for the information that are relevant to our analysis: time stamps, the URLs of the requested pages, the browser configuration of the user. Given the URLs in the user's usage log, one is able to recover the content of the web page. A parser has been developed that retrieves the textual information of the page which then is fed to the actual analysis, featuring pre-processing steps (removal of stop words, stemming etc.), the discovery of the core concepts contained (keyword extraction and disambiguation) and the restructuring of this information to fit the profile ontology representation.

4.2 Information Integration and Exploitation

All information gathered has to be integrated in one single data structure to enable thorough and easy exploitation. This includes on the one hand the results form the above-named evaluation and, on the other hand, information obtained from the analysis' that are already effectuated in-house at Ezakus. The data model has to fulfill a number of crucial demands defined by the commercial setting of its final usage.Based on the demands of their commercial partners, Ezakus has to classify users as belonging to certain market segments. The definition of those segments may differ from one partner to another and over time, the rendering of new segment specifications has to be intuitively. Also, the adaptation of the knowledge structure has to be made automatic.

Other necessities have already been evoked in the description of the project context (Section 2) – the evaluation of user behavior, based on the linkage of the viewed resources has to happen in real time, parallel to his/her actions, even when encountering vast amounts of data.

4.3 Summary

In the above sections, we provided a short summary of the project context and the tasks to be realized by our system in development. Given the ambitious goal

of the project and the constraints imposed by the context of the commercial partner, we need to adopt a highly flexible, structured and intuitive data model that, additionally is able to answer complex requests in real-time. In consequence, we took the choice to adopt an ontological structure for the representation of information throughout the complete profiling process.

Ontologies and their formal languages allow the expressive description of the informational content within all processing steps. Not only do they provide the means to connect the concepts to clearly defined external knowledge sources, but also to specify customized relations between them. Those may extend the commonly adopted taxonomic relations by rich descriptions adapted to the specific context. Therefore, Semantic Web resources are integrated in the analysis process as pre-existing, community-based knowledge resources. They provide the evidence to disambiguate terms and background knowledge to help the deduction of concept's relations. These resources are mainly obtained through the integration of Linked Open Data sets, such as DBPedia[2], Freebase[3]. By using state-of-the-art storage technology for the management of semantic data, we will still come up to the demands concerning real-time and big data processing.

5 The Ontology

Building an ontology is an incremental process of transforming implicit domain knowledge to its explicit representation. Thus, as a first step, we aimed to gain a thorough understanding of Ezakus' working context, the concepts involved and their relations. The result is a customized application ontology, comprising the specification of entities in the ecosystem, such as Ezakus' partners and their affiliated domains and websites, but also abstract concepts such as the user profile itself, the keywords used in the analysis process and so on.

All analysis centers around a BID ("Browser ID"), identifying a user when surfing on a partner website. Attributed to this BID are "hits", the minimal entities of user behavior, each one carrying a time stamp, the URL of the website visited and basic information about the user's browser. Consecutive hits can be assembled to a "session", a set of hits that has been effectuated in no larger temporal distance than thirty minutes. Figure 2 shows a visualization of the upper-level classes of the ontology.

From the context of Ezakus, we include supplemental information about the web pages. These information build the base for our system's analysis. The contents of the web page are extracted and processed. The resulting keywords and, if need be, means to their disambiguation, are fed to the ontology. Known keywords can be directly related to the defined categories. For unknown instances, the associated topics relations and their weighting has to be computed. Combined with the results of Ezakus' internal application of enhanced statistical analysis and machine learning, this gives the final user interest profile.

[2] http://dbpedia.org/
[3] http://www.freebase.com/

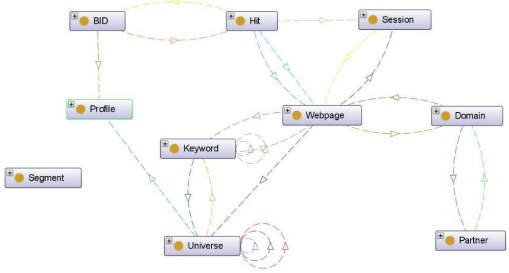

Fig. 2. The high-level concepts of the customised ontology (visualised using the OntoGraf-plugin for Protégé)

The ontology has been modeled in Protégé and then published on a OWL Triple store for its automatic population. Figure 3 shows the an example of the usage of the inference engine for customer segmentation. The segment "Sporty-Mom" has been defined as a user that (a) is female, (b) belongs to a certain age group, (c) lives in a household with children (all the latter summed up by using a pre-defined segment "Mother") and (d) being interested in articles treating sports-related topics. Statements added by the inference engine are marked with a yellow overlay – thus, the affiliation of the instance "user1" to the segment class has been computed automatically.

Fig. 3. Example of automatic inference for the segment "SportyMom" (signifying a person that was identified female, with children and interested in topic "Sport")

Ezakus objective of targeting web users for commercial usage leads us to add another abstraction layer. Marketing segments are a classification of customers, specified by a set of rules applied to their profile attributes. Those rules are included in the ontology to enable the inference engine to automatically deduct a user profiles affiliation to a defined segment.

6 Prototype

A prototype of the system has been developed during the recent months. It provides a proof of concept for our approach and validates the proposed workflow.

The core element of the prototype is the above described enterprise ontology that has been in close collaboration. Even though ontology engineering is an iterative process, it is a first exact representation of the knowledge generated and processed in the context of Ezakus. Web pages are automatically parsed and textual elements extracted. After pre-processing using a stemming algorithm and a part-of-speech tagger [11], a basic tf/idf-implementation is used to discover pertinent keywords to represent the core topics covered.

Ezakus already employs a topic-oriented categorization in their internal processes, and it seems reasonable to stay as close to this existing taxonomy as possible in all stages of the implementation. For the moment, all categories are attributed with a list of keywords that suggest a close relation between a resource and the category in question. These word clouds have been used to compute the similarity between a textual resource and the category, employing a word vector-based approach. To use the information available in already existing external knowledge sources, we provide a mapping between those customized categories and a community-accepted classification scheme, such as the Open Directory Project[4]. Apart from enabling the exploitation of those structured external resources, this will also facilitate evaluation of our approach by the scientific community. Based on this computation, every entering web resource is classified to one of the pre-defined categories. Using the distance value that results from the vector comparison, the membership can be quantified to a certain degree.

All information available about the resource, such as its URL, keywords, assigned categories is entered to the ontology. Additionally, information from Ezakus' internal working processes are integrated, comprising results of machine learning approaches and basic values calculated directly from the user log files. As a matter of fact, the currently used market segments have been defined in a rule-like fashion so far. The known definitions are included in the ontology as classes to enable inference on a class- and instance-level.

For the exploitation of the integrated information, a server has been set up using OWLIM and allowing queries using Sesame. Benchmark tests gave response times of about 5ms for the insertion of a new triplet to the knowledge base and of about 40ms to answer a posed query. Those values conform with the constraints posed by Ezakus at this day.

7 Scientific Contribution

The ontological base structure allows to not only measure the similarity of two sets of words, but to consider additional features of evaluation, especially features of semantic nature. Those enhanced distance evaluations will confront us with a number of issues to solve in our research work:

Text analysis. As the analysis of textual resources has been a key issue to numerous research domains (e.g. information retrieval, cross-language applications),

[4] `http://www.dmoz.org/`

there is a rich body of research contributing to the resolution of problem tasks. However, most tools have been developed for English language while the support for French is still limited. The platform GATE[5] offers utilities and frameworks for the development of customized language processing solutions – and may thus offer a good starting point to enhance existing solutions for French language.

Different Degrees of Structure: The integration of heterogeneous knowledge sources involves not only to handle different semantics, but also varying degrees of specificity of those semantics. A certain analyzed document is represented by a few connected concepts, whereas a carefully constructed word cloud associated to a category may comprise several hundred, highly connected concepts.

Interpretation of Relation Types: In the internal storage as well as in the external sources, relations connecting the concepts can have different types. There is comparably few research work on how to deal with the interpretation of those types for distance computation. Some approaches do already use the synonym-sets from WordNet to extend keyword sets – but how to proceed, for example, when coming across a link that states a certain degree of similarity or, in contrary, the fact that two terms are antonyms? Depending on the application context, those connection types may bear rich information. We will examine their influence on performance in the upcoming stages of the project.

8 Conclusion and Future Work

We describe a novel architecture for user profiling from implicit data. Using navigational data of web users, we adopt natural language processing methods for the analysis of the underlying web resources. If need be, we use external knowledge sources for disambiguation and relation discovery. One important research issue is the adaptation of all algorithms and background resources to French language. All process steps have to be realised respecting the constraints of the industrial partner, in particular concerning the issues of working in real-time and on large-scale data.

After the correct construction of the profile, its maintenance becomes the next issue to tackle. Updates have to be incremental, as it is not possible to store and re-analyze every information when a user re-enters the vision of the system. We adopted a weighting scheme within the ontology to model the degree of affiliation of a customer to a certain segment. To capture the evolution related to short- and long-term interests, an appropriate update mechanism has to be developed. Promising results come from approaches using Spreading Activation for that objective (e.g. [16]).

Acknowledgements. The research work presented in this paper is supported by Ezakus Labs HQ (www.ezakus.com).

[5] `http://gate.ac.uk/biz/usps.html`

References

1. Calegari, S., Pasi, G.: Personal ontologies: Generation of user profiles based on the {YAGO} ontology. Information Processing & Management 49(3), 640–658 (2013)
2. Degemmis, M., Lops, P., Semeraro, G.: A content-collaborative recommender that exploits wordnet-based user profiles for neighborhood formation. User Modeling and User-Adapted Interaction 17(3), 217–255 (2007)
3. Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A.: User profiles for personalized information access. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) Adaptive Web 2007. LNCS, vol. 4321, pp. 54–89. Springer, Heidelberg (2007)
4. Gentili, G., Micarelli, A., Sciarrone, F.: Infoweb: An adaptive information filtering system for the cultural heritage domain. Applied Artificial Intelligence 17(8-9), 715–744 (2003)
5. Ghorab, M.R., Zhou, D., O'Connor, A., Wade, V.: Personalised information retrieval: survey and classification. In: User Modeling and User-Adapted Interaction, pp. 1–63 (2012)
6. Labrou, Y., Finin, T.: Yahoo! as an ontology: using yahoo! categories to describe documents. In: Proceedings of the Eighth International Conference on Information and Knowledge Management, pp. 180–187. ACM (1999)
7. Ma, Z., Pant, G., Sheng, O.R.L.: Interest-based personalized search. ACM Transactions on Information Systems (TOIS) 25(1), 5 (2007)
8. Magnini, B., Strapparava, C.: Improving user modelling with content-based techniques. In: Bauer, M., Gmytrasiewicz, P.J., Vassileva, J. (eds.) UM 2001. LNCS (LNAI), vol. 2109, pp. 74–83. Springer, Heidelberg (2001)
9. Mizoguchi, R.: Part 3: Advanced course of ontological engineering. New Generation Computing 22(2), 193–220 (2004)
10. Qiu, F., Cho, J.: Automatic identification of user interest for personalized search. In: Proceedings of the 15th International Conference on World Wide Web, pp. 727–736. ACM (2006)
11. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, vol. 12, pp. 44–49 (1994)
12. Semeraro, G., Lops, P., Degemmis, M.: Wordnet-based user profiles for neighborhood formation in hybrid recommender systems. In: Fifth International Conference on Hybrid Intelligent Systems, HIS 2005, p. 6. IEEE (2005)
13. Shen, X., Tan, B., Zhai, C.: Implicit user modeling for personalized search. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp. 824–831. ACM (2005)
14. Sieg, A., Mobasher, B., Burke, R.: Ontological user profiles for representing context in web search. In: 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology Workshops, pp. 91–94. IEEE (2007)
15. Stamou, S., Ntoulas, A.: Search personalization through query and page topical analysis. User Modeling and User-Adapted Interaction 19(1-2), 5–33 (2009)
16. Su, Z., Yan, J., Ling, H., Chen, H.: Research on personalized recommendation algorithm based on ontological user interest model. Journal of Computational Information Systems 8(1), 169–181 (2012)