Towards using DBpedia for building user identities

Agata Filipowska and Jacek Małyszko

Poznan University of Economics
Faculty of Informatics and Electronic Economy
Department of Information Systems
Al. Niepodleglosci 10, 61-875 Poznan, Poland
{firstname.lastname}@kie.ue.poznan.pl,
http://www.kie.ue.poznan.pl

Abstract. Internet offers a number of various services that to maximise the user experience apply different personalisation techniques. An important resource of every personalisation method is a user profile. The more information on the user is available in such profile, the better. Therefore, together with maturing of these mechanisms, the notion of identity emerged. The identity exceeds the user profile with information that is more detailed or enables benefiting from additional functionalities. The information stored within an identity needs to be understandable for different services to be easily reused. This can be achieved using the DBpedia.
The goal of the article is to describe the design of a method that potentially enables providing data to build the user identity, based on his behaviour on the Web. The method is elaborated as well as an example of application is presented.

Keywords: DBpedia, Wikipedia, information extraction, identity

1 Introduction

Most users leave a significant amount of information about themselves on the Web. They abandon their anonymity freely (sometimes unconsciously), in order to stay connected with their friends on the social networking sites, communicate with their government or build their reputation [1], [9]. Also, the service providers want to learn detailed characteristics of their users by using different profiling practices [5], in order to provide a better service and preserve their customers. As a result of these trends, a problem emerged of how the users should establish and manage their presence on the Web, namely their digital identities. This issue is being researched for many years now [5].

One of the major challenges concerning the identity management systems is creation and maintenance of many perspectives on users identity, called virtual identities, most preferably without explicit actions of the user. Virtual identity is understood as a collection of topics concerning specific interest of a user. In

this paper we present a method that enables automatic identification of such topics using Wikipedia and information extraction techniques. The method is developed for the Polish language. It utilizes Wikipedia concepts but can easily be extended to DBpedia resources. As there is no Polish DBpedia yet, this will not be covered by this article. However, the work on Polish DBpedia is ongoing and this will be addressed in the future.

The remainder of the paper is structured as follows. Section 2 is devoted to a short summary of virtual identity definitions. In the next section, we indicate current projects and existing approaches that raise the issue of users virtual identities and provide solutions in this area. Section 4 describes the method proposed to identify concepts building the users identity. Finally, in Section 5 we focus on a Use Case demonstrating the application of the method. The article concludes with the final remarks.

2 Definition of Identity

The concept of an identity has been adopted by Information Science as a formal representation of knowledge about a certain person, or any other (digital or real-world) subject. Concerning an identity of a person, it is understood as a set of attributes (permanent or temporary) characterizing a person [13], that is required by providers of services that the person uses [8]. Obviously, a virtual identity cannot capture all characteristics of a person; it is therefore only a partial representation of a subject [13], [14]. Traditionally, an identity is considered as a permanent entity, persisted in a kind of a datastore in order to be accessible many times for a long period of time. However, it can be also understood as something created on-the-fly and used (attached to a person) only during a single session, while a user performs certain tasks or when a particular transaction is performed [7], [13].

More generally, a virtual identity can be defined as a digital representation of a set of claims made by one party about itself or another digital subject [3]. A natural person (a human being) is one example of such entity; other example is a whole organization (i.e. juridical person) [14]. An identity can either be used in a single environment (for example, in a single system or company), or used in many different environments, for example across organizational boundaries. At the same time, different information about every entity is exchanged in different contexts; for example, different user characteristics are needed in e-banking portals and in movie recommender systems. We can therefore either say, that a virtual identity is just one set of claims about a digital subject and for any given digital subject there will typically exist many virtual identities [7], or that each subject has only one identity, but such identity has multiple facets, that are used depending on the context [13].

The identity of a digital subject can be established by combining both the real-world attributes (for example name, address, social security number, physical traits, etc.) and the digital ones (such as passwords, access rights, biometrics, type of encoding, network address and so on) [6]. The information stored in an

identity can be used either for the authentication purposes (its goal is to ensure, that a certain person is indeed what he or she claims to be), or as the attribute information (representing the details about the person) [14]. A set of processes relating to the disclosure of the information about the person and usage of this information is called identification [13].

For the requirements of the "Ego - Virtual identity" (Ego) project[1], presented in the paper, the identity is understood as an information structure describing the information needs of a user. This structure is grounded in the Wikipedia concepts' graph to ease its maintenance and assure usefulness while personalizing information content, especially from the information needs evolution point of view. The future work concerns extending the method towards DBpedia resources.

3 Related work

In the following sections, we present the state of the art analysis of the identity management systems on the Internet in terms of the business goals, that they pursue and the functionalities, that they provide. We identify the most important projects and solutions in the area of identity management systems, that may benefit from the approach we suggest. The main projects that we concentrated on are following: FIDIS[2], SWIFT[3], PICOS[4], PRIME[5], STORK[6], ProjectVRM[7]. Moreover, there exist also frequently updated lists of identity-related efforts[8].

In addition to the above-mentioned projects, a number of already implemented solutions were analyzed. These solutions however mainly focus on the authorisation aspect, leaving behind the notion of user representation e.g. the OpenID protocol describes a user with a limited set of attributes only [11]. Similar, authorisation focused, approaches are e.g. [12], [2], [4]. An interesting, and comparable to ours effort is WebID [15] that uses FOAF vocabulary to describe a user.

Some of the solutions are widely used in business, for example the OAuth protocol [9] or various OpenID implementations, while some of them are at earlier stages of development and adoption, e.g. WebID and Higgins.

Finally, its also very important to indicate, that several organizations have emerged and aim at consolidating and coordinating efforts in the area, of which

[1] http://kie.ue.poznan.pl/en/project/ego-virtual-identity
[2] http://www.fidis.net/
[3] http://www.ist-swift.org/
[4] http://www.picos-project.eu/
[5] https://www.prime-project.eu/
[6] https://www.eid-stork.eu/
[7] http://projectvrm.org/
[8] For example: http://personaldataecosystem.org/2011/06/startup/, http://blogs.law.harvard.edu/vrm/development/, accessed on 15/10/2013
[9] It is used for example by Facebook, Google and Last.FM

the most important are probably Kantara Initiative[10], Identity Commons[11] and formerly Liberty Alliance[12].

It can be easily noticed, that the concept of virtual identities is heavily studied. Nevertheless, as it is a wide field to investigate, different areas of virtual identity creation, maintenance and usage can be explored by different projects. To the best of our knowledge, the approach focusing on automatic creation of users virtual identity that links experience from the fields of information extraction and Wikipedia does not exist.

4 Approach and methods used

This section presents details of the approach we apply to create the identity of a user. The phases of creating the users virtual identity are as follows:

Phase 1: Tracing user behavior. The first step towards building a user's identity concerns identification of topics of user's interest. Of course, these topics can be entered manually by a user (a so-called explicit user modeling [17]), but the identity management systems usually provide additional functionalities to make the whole process more effective.

There is a lot of information about a user even before she or he starts using a given identity management system. Such information is often spread across multiple domains such as web portals, social networking sites, etc. Therefore, the identity management systems can try to somehow import and aggregate information about the user from such sources automatically. To make that feasible, the user's data export mechanisms must be made available by owners of such systems. An example of such initiatives are Data Liberation Front[13] and Data Portability Project[14].

We build the identity of a user based on a wide range of his activities on the Web. Our goal is to engage the service providers in this process, as discussed in [16]. At the current stage of the experiment, we focus on building user's identity based on analysis of the Web pages the user visited. To that end, we have implemented a Web browser plug-in, which a user has to install and have it enabled while browsing. The plug-in extracts (structural, XSLT extraction) the main content of the website and commits it on the server.

Phase 2: Analysis of the visited Web sites. The content that is uploaded to the server is analyzed using the lexical extraction module to identify the differentiating phrases and assign a topic. For the list of topics that are the most representative for the whole content of the network, we chose the Wikipedia categories and concepts list.

The extracted content of the website is analyzed using NLP to identify named entities, cross references, etc. and as a result provide a set of words (surfaces

[10] http://kantarainitiative.org/
[11] http://www.identitycommons.net/
[12] http://projectliberty.org/
[13] http://www.dataliberation.org
[14] http://dataportability.org/

existing in the text, further being referred to as phrases), that will be subject to further processing. What is important, that the approach works for the Polish language and is contextual.

Phase 3: Building a representation of a website for the needs of the identity building. The most crucial step, from the point of view of this paper, as well as for the user acceptance of the system being developed, is indication of a topic, the website mentions. This is done in the following steps.

Firstly (in the preparatory phase), all Wikipedia pages are processed in order to identify concepts (Wikilinks) that appear on these pages in order to learn a phrases-concepts mapping, similarly as it was done by [10]. This process is repeated periodically. Thus, we have obtained 5.150.143 phrase – concept mappings. This mapping is ambiguous, as many phrases may point to many different Wikipedia concepts (on average, each phrase points to 1.21 concepts, but there are some phrases that are mapped to up to 4000 concepts). Still, based on that for each phrase we are able to retrieve a list of candidate concepts.

The method of indication of a topic of a website assigns to each phrase from the text (f) concepts from Wikipedia ($c1$ – $c6$ in Figure 1) obtained as indicated in the previous paragraph. Then, for these concepts ($c1$ – $c6$), the upper level categories of concepts are indicated ($c11$ – $c51$). The Wikipedia category structure enables to build a whole tree over the initial concepts that were assigned, e.g. for concept *Peter Higgs*, based on the Polish Wikipedia structure, we retrieve categories such as *Scottish Physicists*, *Born in 1929*, etc. Currently, we use only three levels within the tree (experimentally evaluated). Then, using the bottom-up propagation method the first-level concepts (mapped from phrases extracted from the website content) vote for the upper level concepts. The bottom up propagation measure combines five frequencies:

- The number of times a phrase from the article text refers to a concept from the Wikipedia.
- The number of times a phrase (surface form) appears in the Wikipedia.
- The number of times a given concept is referenced within the Wikipedia.
- The frequency of a word in the language (in our case the Polish language).
- The number of sub-concepts of a concept.

As a result of the bottom-up propagation, we identify a concept (not necessarily the top-level one), that is the most probable topic of the website. Afterwards, the phrase from the website being the most strongly connected with the concept assigned as a topic, is removed from the initial list of phrases and the procedure is repeated for the remaining phrases. While experimenting, we identified that for most of the articles three iterations are enough to provide the most meaningful concepts describing the website's topic.

These concepts are then mapped on the user's virtual identity. Each new package of topics, changes the initial identity. The weights assigned to different topics within the identity, reflect also maturing in time. Also, user may support this process by manually extending the list of automatically assigned categories.

The user identity created by the system, may be then further used for the needs of personalization of websites visited by the user. The Ego system is to pro-

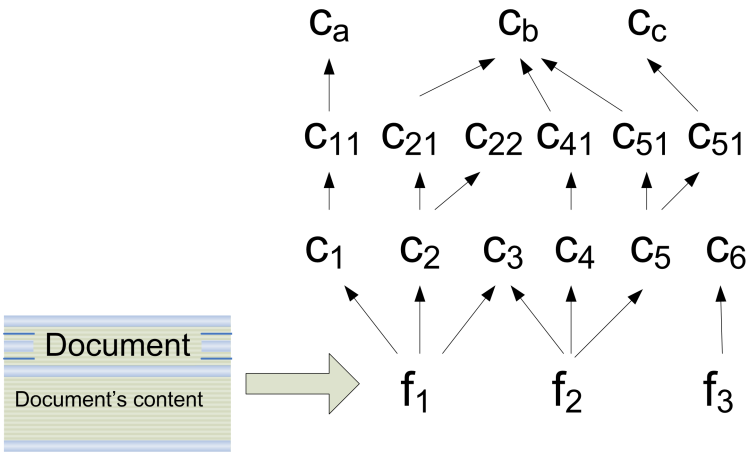

Fig. 1. The multi-layer representation of the article: phrases extracted (f) and Wikipedia concept hierarchy (c).

vide a number of functionalities enabling for sharing and encrypting the identity, authorizing a service provider as well as enabling user to manage the identity and to access it [16].

5 Use Case-based Validation

The presented approach is about to be validated with the real users, who committed to use Ego for a certain period of time and share their experiences. Up till now, the Use Case-based validation has been performed. For the sake of clarity, we present details based on one news article only. The article concerns the Noble Prize Winner Peter Higgs (it is in Polish and is available at Polskie Radio website[15].

The content of the article was extracted and loaded in the database as a logical document (this concerns the topic and the content of the article; menus, comments etc. are not further analysed). Then, the lexical extraction rules extracted 44 different phrases from the article e.g. uroczystości (celebration), professor, etc., out of which 31 were mapped on Wikipedia phrases.

For these Wikipedia phrases, 2052 Wikipedia concepts were retrieved (identified by different URLs) including three upper levels (2052 is a total number of concepts in the tree initially representing the topic of the article). The most frequent concepts in the first level mapping were e.g. fizyka (physics), konferencje miedzynarodowe (international conferences), mechanika kwantowa (quantum physics), II wojna światowa (second world war).

[15] `http://www.polskieradio.pl/23/266/Artykul/951564, Profesor-Higgs-zapadl-sie-pod-ziemie-`

Then, the relations between different concept categories were exploited using the bottom up propagation method. After applying the method, the following concepts were identified as the most descriptive for the article (in the order of importance):

- Urodzeni w XX wieku (born in XX century),
- Popularność (Popularity),
- Higgs,
- Szkolnictwo wyższe (Higher education),
- Nauki przyrodnicze (Natural sciences).

These concepts may be then further mapped on the Wikipedia category structure graph representing the users identity, but this issue is beyond the scope of this paper.

6 Conclusions and future work

The goal of this paper was to present a method that enables for identification of topics that are of user's interest using Wikipedia and information extraction techniques, and based on the behavior of a user on the Web. Starting from a general summary of the Virtual Identity definitions, we presented a method that may be used in order to create user identities using Wikipedia. We also demonstrated an application scenario.

The future work will especially be devoted to tuning of mechanisms developed as well as carrying out an extensive validation of the approach with the real users. The major issue that needs additional research is the bottom-up propagation method that should eliminate concepts being pointed from the multiple websites such as e.g. born in XX century.

Further research will also concern changing the Wikipedia to the DBpedia to allow for an extensive reasoning. This could also offer additional functionalities to an identity management system and service providers that will benefit from it. However, the work on the Polish DBpedia is still the ongoing effort.

Acknowledgments. The work published in this article was supported by the Polish Ministry of Science and Higher Education (decision no. 0987/R/H03/2010/10), upon contract with the Polish National Centre of Research and Development (NCBiR) (contract no. NR11-0037-10/2011) on the project titled: "Ego – Virtual Identity" (http://kie.ue.poznan.pl/en/project/ego-virtual-identity).

References

1. M. Bernstein, A. Monroy-Hernandez, D. Harry, P. Andre, K. Panovich, and G. Vargas. An analysis of anonymity and ephemerality in a large online community. In *5th International Conference on Weblogs and Social Media (ICWSM)*, Menlo Park, 2011. The AAAI Press.
2. C. Burton. The Information Card Ecosystem: The Foundamental Leap from Cookies & Passwords to Cards & Selectors. `http://wiki.informationcard.net/files/icf-information-card-ecosystem-white-paper.pdf`, 2011.

3. K. Cameron. The laws of identity. http://msdn.microsoft.com/en-us/library/ms996456.aspx, 2005.
4. E. Hammer-Lahav. The OAuth 1.0 Protocol. *Internet Engineering Task Force (IETF)*, 2011.
5. M. Hildebrandt and J. Backhouse. D7.2: Descriptive analysis and inventory of profiling practices. FIDIS (Future of Identity in Information Society) Project Deliverable. http://www.cosic.esat.kuleuven.be/publications/article-827.pdf, 2005.
6. J. Hodges, R. Philpott, and E. Maler. Glossary for the OASIS Security Assertion Markup Language (SAML) V2.0. http://docs.oasis-open.org/security/saml/v2.0/saml-glossary-2.0-os.pdf, 2005.
7. Identity Commons. Identity landscape. http://wiki.idcommons.net/Identity_Landscape, 2012.
8. Kantara Initiative Wiki. Consumer Identity Workgroup – scenarios, use cases, & definitions v0.3 . http://kantarainitiative.org/confluence/pages/viewpage.action?pageId=38371527, 2012.
9. R. Leenes, J. Schallaböck, and M. Hansen. Prime white paper. http://security.future-internet.eu/images/2/27/Prime_White.pdf, 2008.
10. D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM, 2008.
11. OpenID Community. OpenID Authentication 2.0 - Final. http://openid.net/specs/openid-authentication-2_0.html, 2007.
12. N. Ragouzis, J. Hughes, R. Philpott, and E. Maler. Security Assertion Markup Language (SAML) V2.0 Technical Overview. http://www.oasis-open.org/committees/documents.php?wg_abbrev=security, 2006.
13. K. Rannenberg, D. Royer, and A. Deuker. *The Future of Identity in the Information Society: Challenges and Opportunities*. Springer, 1st edition, 2009.
14. M. Rundle, E. Maler, A. Nadalin, D. Reed, and D. Thibeau. The Open Identity Trust Framework (OITF) Model (White paper). http://openidentityexchange.org/sites/default/files/the-open-identity-trust-framework-model-2010-03.pdf, 2010.
15. H. Story and S. Corlosquet. Web 1.0. Web Identification and Discovery. http://www.w3.org/2005/Incubator/webid/spec/, 2011.
16. D. G. Weckowski and J. Malyszko. On information exchange for virtual identities: Survey and proposal. *IARIA, 2013*, 978-1-61208-249-3:59–64, 2013.
17. P. Zigoris and Y. Zhang. Bayesian adaptive user profiling with explicit & implicit feedback. In *Proc 15th ACM international conference on Information and knowledge management*, pages 397–404, New York, 2006. ACM.