Leveraging Social Data with Semantic Technology

Mauricio Espinoza-Mejía*, Víctor Saquicela*, Kenneth Palacio†, Nuria García Santa‡, Boris Villazón-Terrazas‡
*Department of Computer Science
University of Cuenca, Ecuador
{mauricio.espinoza, victor.saquicela}@ucuenca.edu.ec
†Department of Electrical&Electronic Engineering and Telecommunications
University of Cuenca, Ecuador
kenneth.palacio@ucuenca.edu.ec
‡Intelligent Software Components, iSOCO, Madrid, Spain
{ngarcia, bvillazon}@isoco.com

***Abstract*—Social networks offer the opportunity of communicating and sharing preferences with other users, reflecting in this way the interests, needs, values, and priorities among the members of the community. However, this information has rarely been considered in traditional recommender systems. This paper shows how Semantic Web technologies can be used to model user information and user-generated content in a machine-readable way. The idea proposed in this work is to extract content-based user profiles from the data available in different social networks in order to obtain an image of the user's preferences that can be used to recommend television programs.**

In particular, domain ontologies and specific vocabularies are put together to describe the information that social networks have about their structure and contents. As a proof of concepts, we propose a simple example that shows how to represent the preferences of a user extracted from different social networks.

I. Introduction

Currently, the internet is being used not only to broadcast and share information about business and companies, but it has proliferated as a new source of entertainment and a massive mean of communication for business and end users. A clear example of this situation, has been the creation of online social networks such as Twitter[1], MySpace[2], LinkedIn[3], Google+[4] or Facebook[5], which have come to revolutionize how people coexist in society, since they serve as tools to share information and interact with other users members of a community.

The collaborative participation of internet users facilities making a better use of the information and resources generated. In fact, in some works in the literature (see [1]–[3]) is possible to identify approaches that take advantage of the information generated inside social networks to design a new paradigm of recommender systems. These works show that the integration of social networks can improve the performance of current recommender systems in at least three aspects: i) *the prediction accuracy*, since the additional information about users and their friends obtained from social networks improves the understanding of user behaviors and ratings, ii) the *data sparsity problem* can be alleviated, because of the fact that two people are friends already indicates that they have things in common and so it is no longer necessary to find similar users by measuring their rating similarity, and iii) the *cold-start issue* can be reduced by using information of the preferences of his/her friends, for cases where a user has no past reviews.

To take advantage of this new paradigm of recommender systems, in a previous work [4], this research group designed a personalized recommendation system, where, the data to model (the user profile and the television programming in our case) were described on terms of semantic concepts defined in ontologies and enriched with information of external sources (social networks such as Facebook or Twitter, movies database such as IMDb[6], repositories of programming guides such as XMLT[7], and SPARQL Endpoints with data of television programs such as BBC[8], etc.). This paper, documents how to model the user information, specially that obtained from social networks.

An interesting indicator presented by the National Institute of Statistics and Census (Intituto Nacional de Estadísticas y Censo - INEC[9]) shows that the 28.2% of Ecuadorians who have internet access, use it as a means of communication, i.e., social networks, email, etc. Perhaps, the most relevant data for the objectives of this work is that the 6.81% of Ecuador's population uses social networks (1.081.620 persons[10]), two points more than in the year 2012, and four more than in the year 2011, following the growing of other users worldwide, reaching a level of penetration of 36% approximately (only in Latin America). This last value is very significant for two reasons: i) enables us to leverage the large amount of personal information embedded in social networks[11], and ii) it provides an opportunity to investigate the influence that play these data on the recommender systems.

All the information embedded into social networks could be exploited if the data generated with these applications use representation mechanisms that allow to interconnect both people and objects on the Web in an interoperable, machine-

[1] https://twitter.com/

[2] https://myspace.com/

[3] https://www.linkedin.com/uas/login

[4] https://plus.google.com/

[5] https://www.facebook.com/

[6] http://www.imdb.com/

[7] http://www.xmltv.org/wiki/

[8] http://www.bbc.co.uk/programmes

[9] http://www.ecuadorencifras.gob.ec/

[10] Source: National Survey of Employment, Unemployment and Underemployment, ENEMDU 2013.

[11] Projecting these data, for the year 2017, when it is intended migrate fully to the ecuadorian digital television standard, at least 52% of the population of the region will use social networks.

understandable, and extensible way. The Semantic Web provides models to capture and integrate the data of different social networks. However, as described in [5], in order to enable a user's access to multiple sites, portability between social media sites is required in terms of (1) identification, personal profiles and friend networks and (2) user's content expressed on each site, whether it is about pictures, contacts, likes, or any type of information. In [6], [7] a possible solution to the first requirement (1) was provided, whereas to fulfil the goals of the requirement (2), the use specific vocabularies that allows modeling the interaction of the user with social networks is proposed.

The remainder of this paper is organized as follows: First, in Section II are discussed in more detail the challenges to overcome in the current social networks. Then, the specific vocabularies used to represent on-line activities of different social networks are reviewed in the Section III. Finally, before concluding, Section IV discuss the application of semantic technologies to enhance current social media sites with a sample of the digital TV domain.

II. Understanding Social Networks

A Social Network (SN) is commonly intended as a communication platform that promotes collaborative and participative behaviour of profiled human users [8]. Depending on the scope, coverage and use, social networks can be categorized into two groups: *internal social networks* and *external social networks*. In this paper, mainly the second group (*external social networks*) is considered, because these networks present publicly on the Web the user profiles facilitating information retrieving.

In general, online social networking sites offer services that are valuable in different areas such as bussiness, education, etc; however, these sites present problems and challenges that must be resolved to exploit all the information generated. Fernández, et al., has recently compiled some of such problems in [9]. A short summary of these problems is described below:

- *Heterogeneity in the data and formats used*. Although, current social networks provide an easy access to user profile data through dedicated APIs, these functions do not provide accurate information on the schemes of response This situation complicates the access to information and the possible integration of data obtained from different social networks. Different options have been proposed in the literature to alleviate this problem: i) to develop portable analysis models, ii) to allow users to access their data uniformly across social networks, and iii) to allow automatic data portability from one social network to another one.
- *Many isolated communities of users with their data*. In the current organizational "structure" of the online social networks, a single user might register in several different social networking sites all for different purposes. User information is scattered over these different social networks and stored at the sites own data silos. This issue requires an effective recovery of all available data about a person from various social networks and the integration of these data into a single profile.
- *Inadequate representation of the data*. In order to map the knowledge exchanged by the users of social networks and promote interoperability, a good balance in the standardization of the manifold ways of describing content on these applications should be found. Rather than requiring proprietary APIs to access this data from each service, uniform representation mechanisms are needed to represent and interconnect people and objects on the sites in an interoperable, extensible way. Therefore, the goal is to find ways in which machines can interpret social media contents, to improve the processes of information retrieval and recommendation.

The authors of this approach have explored some of the above problems. We have investigated the use of domain ontologies as an enhanced modeling ground to incorporate meaning to the data captured from social networks, so that the user profile can easily be exploited by a machine [7]. This approach offers a unique entry point for personal data across different social media sites. The following is a brief summary of the main features of this approach:

- A manual identification process is used to detect different profiles belonging to the same individual. This process eliminates the need for an algorithm aimed to solve the problem of entity recognition for profiles found in different social networks. The HybridAuth library[12] was used to implement the user authentication process on different social networks. After authentication, HybridAuth provides the connected user profiles in a rich, simple and standardized structure across all social networks APIs.
- The ontological model requirements were identified by instantiating the NeOn methodology guide for requirements elicitation [10]. To identify the functional requirements, the ontology development team mainly used the technique of writing the requirements in natural language in the form of the so-called competency questions (CQs). CQs are natural language questions that the ontology to be built should be able to answer. Moreover, simple heuristic techniques of information extraction were adopted to extract the pre-glossary of terms of the ontology from the list of CQs and their answers.
- In order to build the ontology, a top-down approach based on the scenario six of the NeOn methodology was adopted. This scenario proposes the reuse, fusion, and re-engineering of existing ontological resources. After performing the search, comparison, selection, customization, and integration processes proposed in the methodology, nine ontological models were merged and modified through a process of re-engineering.
- Social annotations are preprocessed to facilitate matching unstructured data from some social networks (e.g. Tweets) with concepts of the ontology. This pre-processing process involves morphological and semantic transformations of the data. Different natural

[12] http://hybridauth.sourceforge.net/

language processing tools (e.g. Tokenizer[13], Gate[14] and Freeling[15]) were used for the implementation of the activities that allow to clean and separate the input text into words. To discover the correct meaning of a term depending on the context, a disambiguation mechanism based on the overall semantic knowledge base and multi-domain, -Wikipedia[16]-, was used.

- A manual matching process is used to perform the annotation of an entity recovered from a social network entry to a concept in the ontology. Data of each retrieved profile is then exported to RDF[17] and combined using owl properties such as: `sameAs`, which is used to identify that two resources are the same in spite of having different URIs. SPARQL[18] queries are used to store the user profile data into a semantic repository.

The current version of the system supports Facebook, Twitter and Google+ as external sources to discover the user TV preferences. However, the inclusion of a new resource into the system, represents a great effort that involves different tasks of programming. To avoid this issue, the inclusion of a semantic model that supports a universal data format and not one specific to any particular site is proposed. The next section describes the most relevant social web vocabularies that could be used to represent all data entries for a given user no matter where they come from.

III. Using Semantics to Model the Social Web

Semantic Web technologies provide standards and models to build a Web of Data, with unified models to represent interlinked data from different sources [11]. We will overview some vocabularies in this section, by distinguishing models created with a general purpose and models created for a particular social application. An extensive review of the ontologies developed for the Social Web can be consulted in [12].

The SIOC Ontology[19] – Semantically-Interlinked Online Communities – is considered as one of the building blocks of the Social Semantic Web. In combination with other domain ontologies SIOC lets developers link user-created content items to other related items, to people (via their associated user accounts), and to topics (using specific "tags" or hierarchical categories) [5]. One of the aims when developing the SIOC ontology was to keep it as simple as possible, yet powerful, so that it could be easily deployed in existing applications [5].

The Tag Ontology[20] was the first RDF-based model for representing tags and tagging actions. This ontology defines the "Tag" and "Tagging" classes with related properties to create the tripartite relationship of tagging. In order to represent the user involved in a tagging action, this ontology relies on the FOAF vocabulary[21]. The Modular Unified Tagging Ontology[22] (MUTO) is an ontology for tagging and folksonomies. It is based on a thorough review of earlier tagging ontologies and unifies core concepts in one consistent schema. It supports different forms of tagging, such as common, semantic, group, private, and automatic tagging, and is easily extensible.

The Social Semantic Cloud of Tags (SCOT) ontology[23] is focused on representing tag clouds and defines ways to describe the use and co-occurrence of tags on a given social platform, allowing one to move his or her tags from one service to another and to share tag clouds with others. SCOT reuses the Tag Ontology as SIOC does to model tags, tagging actions, and tagging clouds. An important aspect of the SCOT model is that it considers the space where the tagging action happened (i.e., the social platform, e.g., Flickr[24] or Delicious[25]). The Meaning of a Tag (MOAT)[26] aims to represent the meaning of tags using URIs of existing domain ontology instances or resources from existing public knowledge bases [13], such as those from the Linking Open Data project [14].

This work uses SIOC to model the structure and activities in online communities, based on the premise that since 2010, SIOC is a core ontology specification at W3C[27]. Moreover, the use of SIOC goes further than popular and mainstream Web 2.0 services, from Enterprise 2.0 information integration[28] to Health Care and Life Sciences discourse representation[29].

IV. Describing TV Preferences using SIOC and Domain Ontologies

This section describes how a combination of the SIOC model and other domains ontologies can be used to capture and serialize the profile of a digital TV user extracted from social networks. Before illustrating, how information about a typical user of social networks can be semantically modeled, some of the classes used in the SIOC vocabulary are briefly analyzed as follows:

Figure 1 describes the main classes of the SIOC core ontology. It provides the basis for defining a user (class `User Account`), the content that he/she produces (class `Item`, property `has_creator`) and the actions of another user on this content (property `has_reply`). SIOC types extend `sioc:Item` to specify different types of resources produced online. A more detailed description of the specifications of the SIOC ontology is out of scope of this paper, however, users interested in this information can refer to [15]

In order to explain the benefits of combining SIOC ontology with others domain semantic models, the use case of a user that tags his TV preferences in social networks is described. Figure 2 illustrates the use of the semantic models used i) to represent personal data, interests, and activities and ii) to identify all content objects produced by a single user on various social networks. In the example, the terms `ont`, `sioc`, and `foaf` represent the prefixes of the TV User Preferences, SIOC and FOAF ontologies, respectively. As pointed out in

[13] http://nlp.stanford.edu/software/tokenizer.shtml

[14] https://gate.ac.uk/

[15] http://nlp.lsi.upc.edu/freeling/

[16] http://www.wikipedia.org/

[17] http://www.w3.org/RDF/

[18] http://www.w3.org/TR/sparql11-overview/

[19] http://rdfs.org/sioc/spec/

[20] http://www.holygoat.co.uk/projects/tags/

[21] http://xmlns.com/foaf/spec/

[22] http://muto.socialtagging.org/core/v1.html

[23] http://www.scot-project.org

[24] https://www.flickr.com/

[25] https://delicious.com/

[26] http://www.moat-project.org

[27] http://www.w3.org/Submission/sioc-spec/

[28] http://www.w3.org/2001/sw/sweo/public/UseCases/EDF/

[29] http://esw.w3.org/topic/HCLSIG/SWANSIOC

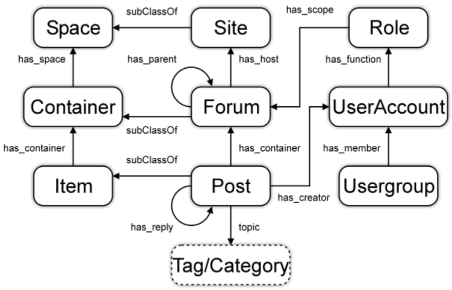

Fig. 1. The SIOC core ontology model [15].

Section II, the TV User Preferences ontology was designed in a previous work.

It should be noted that the great majority of the data that define for example the user's profile *Mauricio* can be extracted from the social networks. In this case the user has two social networking accounts, one in *Facebook* and the other one in *Twitter*. From Facebook the following information is obtained: the locality where he lives (*Cuenca*), his age (*39*), two positive feedbacks (*Two and half Men* and *NBA*) and the fact that one of his friends is *Victor*. Twitter offers important information to the profile; describing that the user follows the page of the Ministry of Environment of Ecuador (*Ambiente Ecuador*) and therefore is likely interested in the environment. Also, in the case of Twitter, the SIOC ontology allows to represent a tweet that the user has entered about the movie "*The Dark Knight*".

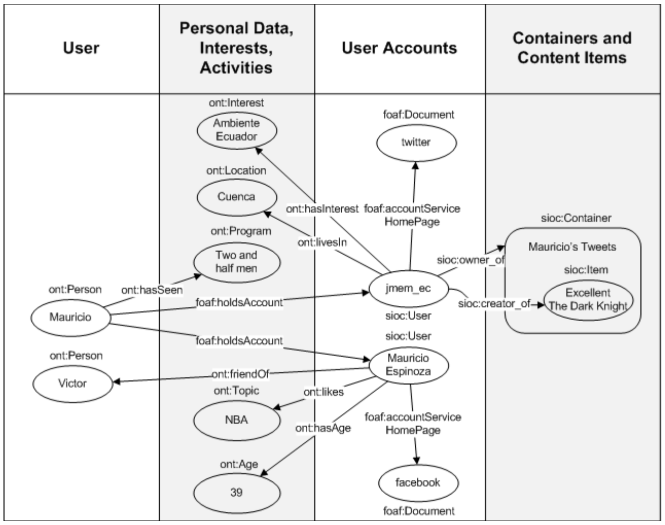

Fig. 2. Modeling User TV Preferences

V. Conclusion

This paper first introduced the challenges to overcome in extracting information from social networks. The main limitation of current social networks is that they are isolated from one another like islands in the sea, acting as closed-world and independent data silos. Then the different ontologies used to represent the Social Semantic Web has been described. In the realm of the Social Semantic Web, ontologies can be then used to represent uniformly the different artifacts produced and shared in social websites: communities, people, documents, tags, etc.

The application of semantic technologies to enhance current social media sites is analyzed. Then by making use of a example in the digital TV context, it has been shown that the combination of domain ontologies and other specific vocabularies contributes to enrich the simple representations of social networks and the content their users share with semantics, in order to fully exploit the wealth of data and interactions on the Web.

Acknowledgment

The research described in this work in progress is part of the project "Application of Semantic Technologies for Reducing the Information Overload in Users of Digital TV" supported by the Central Research Direction of the University of Cuenca (DIUC).

References

[1] J. He, "A social network-based recommender system," Ph.D. dissertation, University of California, Los Angeles, CA, USA, 2010.

[2] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King, "Recommender systems with social regularization," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, ser. WSDM '11. New York, NY, USA: ACM, 2011, pp. 287–296.

[3] E. Davoodi, K. Kianmehr, and M. Afsharchi, "A semantic social network-based expert recommender system," *Applied Intelligence*, vol. 39, no. 1, pp. 1–13, 2013.

[4] V. Saquicela, M. Espinoza, J. Mejía, and B. Villazón-Terrazas, "Reduciendo la sobrecarga de información en usuarios de televisión digital," in *Proceedings of the Workshop on Semantic Web and Linked Data*, Morelia, México, November 2013.

[5] U. Bojars, A. Passant, J. G. Breslin, and S. Decker, "Social network and data portability using semantic web technologies," in *2nd Workshop on Social Aspects of the Web (SAW 2008)*, 2008, pp. 5–19.

[6] M. Espinoza and V. Saquicela, "Modelando los hábitos de consumo televisivo usando tecnología semántica," in *Proceedings of the IX Congreso de Ciencia y Tecnología ESPE 2014*, Quito, Ecuador, Mayo 2014.

[7] M. Espinoza, V. Saquicela, K. Palacio, and H. Alban, "Extracción de preferencias televisivas desde los perfiles de redes sociales," *Revista Politécnica, Escuela Politécnica Nacional*, vol. 34, 2014.

[8] A. Maccioni, "Towards an integrated social semantic web," in *Current Trends in Web Engineering*, ser. Lecture Notes in Computer Science, Q. Sheng and J. Kjeldskov, Eds. Springer International Publishing, 2013, vol. 8295, pp. 207–214.

[9] M. Fernández, V. Rodríguez, A. García-Silva, and O. Corcho, "Social web: Where are the semantics?" Proceedings of the 11th Extended Semantic Web Conference – ESWC 2014 Tutorials, May 2014.

[10] M. Suárez-Figueroa, *NeOn Methodology for Building Ontology Networks: Specification, Scheduling and Reuse*, ser. Dissertationen zur künstlichen Intelligenz. IOS Press, 2012.

[11] M. Challenger, "The ontology and architecture for an academic social network," *IJCSI International Journal of Computer Science Issues*, 2012.

[12] V. D. Breslin J., Passant A., "Social semantic web," *In: Domingue J., Fensel D., Hendler J. (Ed.) Handbook of Semantic Web Technologies*, 2011, springer-Verlag Berlin Heidelberg.

[13] A. Passant, P. Laublet, J. G. Breslin, and S. Decker, "A uri is worth a thousand tags: From tagging to linked data with moat." *Int. J. Semantic Web Inf. Syst.*, vol. 5, no. 3, pp. 71–94, 2009.

[14] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data - the story so far," *Int. J. Semantic Web Inf. Syst.*, vol. 5, no. 3, pp. 1–22, 2009.

[15] U. Bojars, J. Breslin, D. Berrueta, D. Brickley, S. Decker, S. Fernández, C. Görn, A. Harth, T. Heath, K. Idehen, K. Kjernsmo, A. Miles, A. Passant, A. Polleres, L. Polo, and M. Sintek, "Sioc core ontology specification," W3C Member Submission, June 2007.