Collective Knowledge Ontology User Profiling for Twitter

Automatic User Profiling

Paula Peña, Rafael del Hoyo, Jorge Vea-Murguía, Carlos González, Sergio Mayo
Aragon Institute of Technology
C/ María de Luna nº7, Zaragoza, Spain
{ppena,rdelhoyo,jvea,cgonzalez,smayo}@ita.es

***Abstract*—How to model user interests and intentions through user profiling is an important key for providing personalized service on Internet. User profiling can be seen as the inference of user interests, intentions, characteristics, behaviors and preferences. This paper introduces a scalable and automated technique for user profiling by extracting his URLs from publicly available tweets information and using a semantic ontology in which user interests and intentions are characterized. In order to enhance the performance of our method, categorization of websites offered by OpenDNS and DBpedia collective knowledge databases are used to find the interests and intention categories of the user profile ontology. In this context, user profile ontology is populated taking these collective categories and with assertions of individuals, and relationships of interest and intention. As new concepts and relationships are defined and inferred, user profile ontology evolves continuously. Experimental results based on user's tweets confirm strongly that the proposed method improves the automatic acquisition of interests and intentions of a user profile.**

Keywords— Twitter, Social Networks, Ontology, RDF, OWL, NoSQL

I. INTRODUCTION

User profiling is presented as a new tool for Internet services expansion. It can be defined as the inference of user interests, intentions, characteristics, behaviors and preferences. For example, profiling a user's location, buy items or topic interests (which we will focus) enables new services to provide personalized search results, news sites to recommend buy items, and advertisers to serve targeted ads. In order to profile a user, the traditional approaches leverage limited user-centric data (e.g., search log or purchase history), mining values of various user attributes such as demographic characteristics (e.g., age, gender, origin), intentions (looking items to buy, e.g. Smartphone), interests (e.g., sports, travel, entertainment, politics). Also, in new advanced services on Internet such as discovering danger users for a target topic like terrorism or looking for new specific customers, scalable algorithms for mining big data in order to generate user profiles can help. Twitter, LinkedIn, Google+, Facebook and other similar services study the users' posted content and their interactions with others. The real business of these companies is to know about their users to shell advertisements and to improve the quality of experience of them. Beyond this techniques, other methods try to generate online digital footprints [1] for disambiguating the users. A repository of improved user profiles may significantly generate better and new internet business. Ontologies as a formal description and specification of concepts can play an important role in user profiling [2]. They provide a well-defined and constructed method to provide a standard format in order to define user interests.

This work introduces a scalable and automated technique for user profiling by extracting his URLs from publicly available tweets information, enhancing the intelligence of our method and taking the collective categorization of the websites from DBpedia[1] and OpenDNS[2]. A semantic ontology in which user interests and intentions are characterized is used. In fact, we generate a concise but descriptive semantic ontology user profile using Twitter streams. This user profile ontology allows the exact topics of interest that a user has can be easily identified and reasoned. In contrast to bag of word approaches, we generate semantically enhanced user profiles that quantify the users' interests and intentions in a set of specific categories. Ultimately, our profiling method outputs a semantically enhanced user profile that reflects the real user interest.

This paper is organized as follows: Related work on modeling expertise of social media users is presented in Section II. Section III describes the architecture used in our approach and presents a high level description of the algorithm. The user profile ontology used is described in section IV. We discuss our results in section V and highlight implications and of our work. Finally, section VI describes conclusions of our work and discusses ideas for future work.

II. RELATED WORK

Online social network services such as Twitter, LinkedIn, Google+ and Facebook become important means for users to connect with friends and share information. For example, Twitter, a social network for users to follow each other and

[1] DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia (http://dbpedia.org)

[2] OpenDNS cloud websites tagging (http://community.opendns.com/domaintagging/)

978-1-4799-2902-3/13 $31.00 © 2013 IEEE
DOI 10.1109/WI-IAT.2013.62

publish tweets[3], now has almost 500 million active users and generates 50 million tweets daily. On one hand, those services need to "understand" their users better, because old tasks (e.g., targeted ads) now become even more challenging (e.g., serving ads without queries), and new tasks (e.g., recommending "friends") arise in the context of social network. How are people connected on Twitter? Who are the most influential people? What do people talk about? How does information diffuse via retweet? On the other hand, those services generate additional information to leverage, because not only user-centric data (e.g., tweets) is available, but also information from others can be propagated through users' social connections. Profile information including name, age, location in Twitter services, although it can be incomplete (a user may choose not to post bio details) or misleading (a user may choose to list a wrong place). As a micro-blogging site, Twitter is supposed to hold less personal information than sites like Facebook. Despite this, we wondered if it is possible to reconstruct the profiles of Twitter [3] users from only publicly available information on their profile. However, other relevant attributes, such as explicit and implicit interests or political preferences are usually none asked. Different approaches can be obtained in the literature like automatic user classification and profiling politics interests [4]. For example, the problem to discover interest by casting it as a user classification task and leveraging two types of information: user-centric information reflecting the linguistic content of the user's tweets, his social behaviors and likes, and; social graph information in the form of the distribution of the possible target class values for the people connected to the user by a social graph link. Machine learning approach has been used in several occasions like in [5], where is described a general machine learning framework for social media user classification which relies on four general feature classes: user profile, user tweeting behavior, linguistic content of user messages and user social network features. It has been proposed the use of wavelet based on clustering method to group users for discovering regular and consistent behavioral patterns in topical tweeting [6] into different groups that exhibit behavioral similarity. According to [7], it is shed some light on the usefulness of different types of user-related data (tweets, retweets, bio and list data) for making sense of the domain expertise of Twitter users. Also, there are papers working on the identification of the personality of the users [8]. Different works like [9,10] develop tools and services to allow to the end-users to inspect Twitter-based profiles and enables other applications to reuse these profiles. People can overview their personal Twitter activities or profiles of other users to explore the topics those users were concerned with in the past. Different visualization methods (entity-based, topic-based and hashtag-based tag) are used to visualize profiles.

Another topic explored in Social network analysis is to use Twitter as a social virtual sensor [11, 12]. The large number of twitter updates results in numerous reports related to events, including social events such as parties, baseball games, and presidential campaigns. Also, disastrous events such as storms, fires, traffic jams, riots, heavy rainfall, earthquakes or the last bomb attack in Boston can be used as event detector or sensor detector. User profiling or twitter system event detection classifies events that are visible through tweets such as earthquakes, car accidents, terrorist attacks or fires.

Different works exist in the literature to study the topological characteristics of Twitter and its power as a new medium of information sharing [12, 13]. Twitter has explicit social structures among users and can be viewed as a time-series which records the activity volumes of a user at different intervals over an extended time period [12]. In order to identify influences on Twitter, it is possible to rank users by the number of followers or ranking by retweets [14].

This paper aims to find scalable methods to find users with specific interests and intentions to target advertising more tailored to the interests of the people.

III. Proposed Approach

Nowadays, user profiling to model user interests and intentions is an important key in personalized services.

In our domain, a user profile is modeled by ontology based on OWL (*Ontology Web Language*) or RDF (*Resource Description Frame-work*)/XML format. Reference user profiling ontology was initially defined and completed with concepts (or synonyms) extracted from additional sources such as advertisement taxonomies, OpenDNS taxonomy and DBpedia ontology. In addition, our taxonomy defined adheres to the *Friend-Of-A-Friend* (FOAF) ontology.

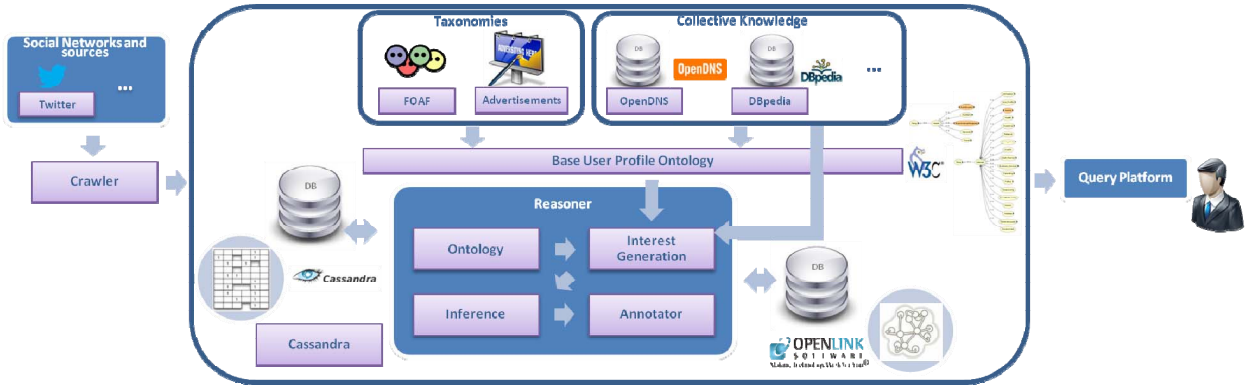

Fig. 1. High level architecture

[3] Twitter stats from http://www.statisticbrain.com/twitter-statistics/

OpenDNS offers a free domain service, *Domain Tagging*, to filter web sites based on several categories. Domains has been tagged into these categories and voted on the accuracy of submitter's tag by community members. Domain Tagging community is composed of tens of thousands of researchers and professional people, from all around the world and including representation from academics, security, technology and various other disciplines. According to the community, a domain can be *awaiting votes* (community is voting on this domain's tags), *approved* (domain is confirmed in the category by the community) or *rejected* (domain does not belong in the category). DBpedia extracts structured data from Wikipedia and make it available as RDF (*Resource Description Framework*). Data can be accessed using an SQL-like query language called SPARQL. OpenDNS and DBpedia knowledge bases play an important role in enhancing the performance of our method. Both cover many domains.

Over the past few years, trends such as concurrency, connectivity, peer-to-peer, mobility and cloud computing have created the need to store large amount of data in distributed databases that provide high availability and scalability. New varieties of non-relational databases, commonly references as NoSQL, have emerged. The loss of flexibility or rigid schemes, the inability to scale data, the high latency or low performance and cost, are some of the major data management problems leading to the adoption of these technologies, nowadays widely used by companies such Amazon or Google. In this context and in order to support the ability to process large amounts of real-time information, NoSQL database become a crucial requirement in our application domain. Non-relational databases *Cassandra*[4] and *Virtuoso*[5] are presented as the most suitable, after a thorough analysis of the current state of NoSQL technologies and solutions.

The high level architecture (Fig. 1) is focused on extracted URLs posted by users in tweets, user profile ontology, OpenDNS, DBpedia and NoSQL databases and Reasoner, seen as an intelligent component that implements the logic and key algorithm to populate user profiles with information about their interests and intention. Besides, it is extensible to other social networks and data sources.

In order to implement the algorithm, we have developed a set of sub-processes that allows adding new functionalities, defining and executing a workflow through an own library for semantic analysis of information inspired in UIMA[6] philosophy, called Moriarty. It works as server processes with BPEL interpreter engine (jBPMN).

URLs contained in tweets can be considered as the seed for our approach. This information is helpful for identifying interest and intention of a user and providing advanced Internet services. In our solution, tweets from users are crawled from Twitter in order to extract URL content, which cover different topics. URLs that people share on Twitter show their interests in specific topics. The ability to classify these URLs allows a first approximation to analyze their interests or their purchase intentions. This information posted by users in tweets is stored in Cassandra NoSQL database. Cassandra is a column oriented distributed storage system, designed to handle very large amount of data spread out across many servers while providing a highly available service with no single point of failure. This allows that the system to be scalable to support from hundred to millions of tweets.

User tweets information gathered in Cassandra database is constantly analyzed in a batch mode. Basically, new information from all users is searched in Cassandra database through *Cassandra* sub-process. User-to-user, with its identification, *OWL2Onto* sub-process performs a search of its existing user profile ontology or a predefined ontological model within Virtuoso NoSQL database of RDF triples. Based on the extracted information, the aim of *InterestGeneration* sub-process is to generate new interest and intention relationships and concepts in user profile ontology. Over these relationships and concepts, *Inference* sub-process deduces new information about interest and intentions by means of a reasoner called Pellet[7] (e.g. if the property *hasIntent* is transitive, and the property relates individual A to individual B, and also individual B to individual C, then it can be inferred that individual A is related to individual C via property). Finally, *Onto2Virtuoso* sub-process allows saving the updated user profile with the new information inferred in Virtuoso NoSQL database. Once generated or updated the ontological profile for all users, *UserProcessedCassandra* sub-process set in the Cassandra database that users have been processed.

We propose a new user profiling algorithm which takes place in several steps and is described as follows:

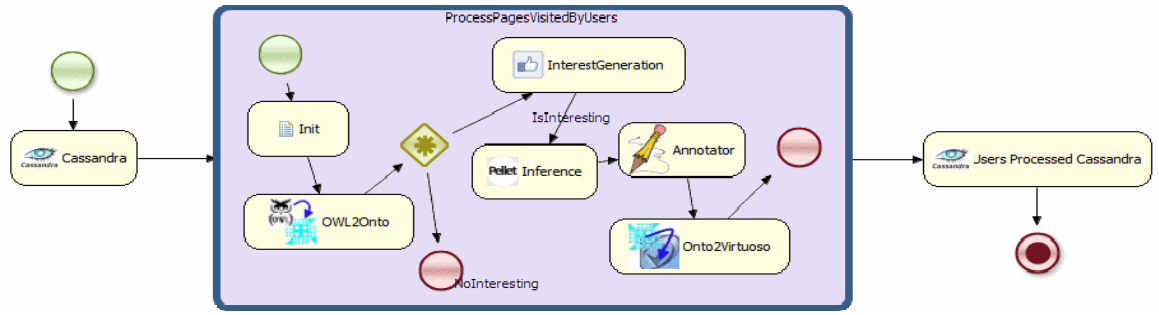

Fig. 2. Workflow and sub-processes

[4] The Apache Cassandra database (http://cassandra.apache.org/)

[5] Virtuoso, a grade multi-model data server (http://virtuoso.openlinksw.com/)

[6] UIMA Project (http://uima.apache.org)

[7] Pellet reasoning server (http://clarkparsia.com/pellet/)

User Profiling Algorithm

Input: url, userId, ontology
Output: user profile ontology with assertions

1. Check if url is interesting (/* not a search engines, a url shorteners or a generalist social networks */);
2. **If** url is interesting **then**
3. **If** url exists in ontology **then**
4. *hasIntent* or *hasInterest* relationship is generated in user profile ontology;
5. **Else**
6. **Search** url **in** [OpenDNS, DBpedia,…]; (/* collective knowledge */)
7. **If** url exits in [OpenDNS, DBpedia,…] **then**
8. **For each** category in [OpenDNS, DBpedia,…] associated to the url
9. *category* in [OpenDNS, DBpedia,…] is added to user profile ontology;
10. *hasIntent* or *hasInterest* relationship is generated in user profile ontology;
11. **Search** sameAs **of** *category* in [OpenDNS, DBpedia,…]
12. **For each** categorySameAs associated to category in [OpenDNS, DBpedia,…]
13. *categorySameAs* is added to user profile ontology like *equivalent class* to *category* in [OpenDNS, DBpedia,…];
14. *hasIntent* or *hasInterest* relationship is generated in user profile ontology;
15. **If** url does not exist in [OpenDNS, DBpedia,…] **then**
16. url is added to user profile ontology into *Unknown* category
17. *hasInterest* relationship is generated in user profile
18. **return** ontology (/* with assertions of relationships and new concepts */);

Fig. 3. User Profiling Algorithm

Our algorithm discriminates between interesting and uninteresting URL. An uninteresting URL includes search engines (such as Google, Yahoo or Bing), URL shorteners (such as Bitly, Goo or Su) and generalist social networks (such as Facebook, Twitter or Tuenti). User profiles are enriched with concepts and topics extracted from OpenDNS and DBpedia knowledge bases enhancing the performance of our approach. An extraction process of Domain Tagging data from OpenDNS has been implemented trough *Web-Harvest*, an Open Source Web Data Extraction tool, in order to obtain URLs categorized by concepts. In addition, with concepts obtained as a result of a query against the DBpedia SPARQL endpoint[8], user profile ontology is populated with RDF assertions of URLs, concepts and relationships of interest and intention. As new concepts and relationships are defined and inferred, user profile ontology keeps alive.

IV. User Profile Ontology

To investigate how domain knowledge can help in the acquisition of user preferences we use ontology. Artificial intelligence literature contains several definition of ontology. Ontology is a term borrowed from philosophy that refers to the science describing the kinds of entities in the world and how they are related. We assume that ontology is a format explicit description of concepts in a particular domain ("class" sometimes called "concepts"), properties of each concept describing various features and attributes of the concept ("slots", sometimes called "roles" or "properties"), and restrictions on slots ("facets", also called "role restrictions").

Ontology together with a set of individuals of classes constitutes a knowledge base. Classes are the focus of most ontologies and describe concepts in a domain. Our base profile ontology is constructed based on standard advertisements taxonomy for user profile, OpenDNS taxonomy, FOAF ontology and DBpedia ontology. Our ontology is always alive, as it may define new classes from the online update with additional concepts from other sources, such as OpenDNS and DBpedia (extensible to other collective knowledge bases). The main classes are described as follows: *Person*: class that contains user identification.

- *URL*: class that includes URLs posted in tweets.
- *Interest:* class that hosts concepts related to relationships with user interests URLs.
- *Intention:* class that holds classes dedicated to relationships involving URLs user purchase intentions.
- *Unknown:* class that contains URLs that do not exists in collective knowledge databases (OpenDNS, DBpedia…).
- *UnknownCategory*: class of categories that do not exist in collective knowledge repositories (OpenDNS, DBpedia…).

The main subclasses defined for intentions (Fig. 4) are *Auctions*, *AutoBuyers*, *Ecommerce/Shopping*, *Services* and *Travel;* the subclasses for interests (Fig. 4) are *Academia*, *Adult_Themes*, *Business_Services*, *Events*, *Government*, *GreenLiving*, *Health*, *Hobbies*, *Humor*, *JobSeekers*, *News/Media*, *Non-profits*, *Parenting*, *Politics*, *Religious*, *Sports*, *TechEnthusiasts* and *TravelEnthusiasts*. The topology of this reference taxonomy built can be showed in Fig. 5.

The basic relationships defined in ontology are "hasInterest" and "hasIntent". Each user classes are related to classes of "Interest" by the relation "hasInterest", and are related to classes of "Intention" by the relation "hasIntention". Evidently, URLs can have a relationship of belonging to one or more concepts from the categories of interest and intentions. The inference is performed to obtain the interests and intentions of each user from de URLs posted in tweets.

V. Results

In order to evaluate our approach for generating user profiles, 18,000 tweets from 8,000 users are been crawled from Twitter in order to extract URL content. In this section, we

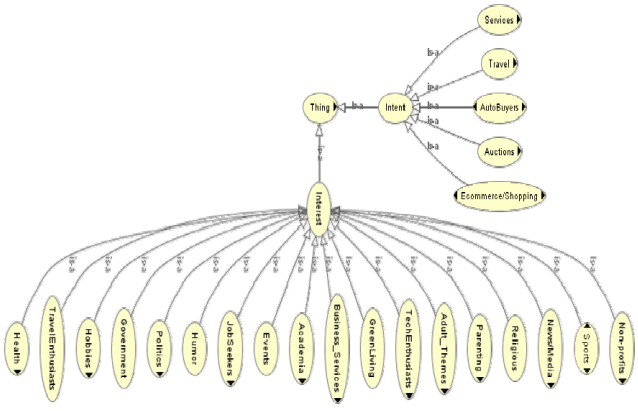

Fig. 4. Intent and Interest first level branch of the reference ontology

[8] Sparql query end point (http://dbpedia.org/sparql)

present results from a sampling of sixteen URLs posted in tweets by nineteen users as an exemplary of a larger experiment conducted. These URLs cover different topics.

Comparing the topology of user profile ontology (our reference or base taxonomy) with the topology of the inferred model, both illustrated in Fig. 5, it can be clearly seen how it has evolved. User profile ontology is populated with information (new concepts and relationships) about their interest and intentions from OpenDNS and DBpedia collective knowledge repositories.

Www.elconfidencial.com tagged into "News/Media" category, *intagram.com* tagged into "Photo sharing" category, *ask.fm* tagged into "Blogs" and "Forums/Message_boards" categories, *apps.facebook.com* tagged into "Games" and "Social networking" categories, *skydrive.live.com* tagged into "File storage" category, *www.ebay.com* tagged into "Auctions" and "Ecommerce/Shopping" categories, *adf.ly* tagged into "Advertising" and "Business Services" categories, *youtu.be* tagged into "Video sharing" and "Software/Technology" categories, *pinterest.com* tagged into "Blogs" and "Photo sharing" categories and *issuu.com* tagged into "New/Media" and "Business Services" categories, are some URLs of our experiment that are only in OpenDNS.

On the other hand, the extracted URLs in tweets *www.avis.com*, *soundcloud.com* and *www.telecinco.es* are examples of URLs that are in OpenDNS and DBpedia.

Filtered the URL *www.avis.com* by OpenDNS, it can be observed that user profile ontology is populated with new concepts related to "Travel" category such as "Car_rental_companies", "Companies_based_in_Detroit,_Michigan", "Franchises", "Transportation_companies_of_the_United_States", "Companies_established_in_1946", "Companies_based_in_Morris_County,_New_Jersey", "Companies_based_in_Nassau_County,_New_York", and their "sameAs" provided by searches in DBpedia knowledge base. In addition, as "Travel" belongs to "Intent" category of reference user profile ontology, "hasIntent" relationships have been generated between user and URL, URL and these categories and consequently, inferred intention relationships between user and sameAs categories enrich the user profile.

In a similar way, the domain *soundcloud.com* is classified such as "Music" category in OpenDNS. Searches in DBpedia allow complete the user profile ontology with new concepts such as "Social_networking_services", "Streaming", "Streaming_music_services", "Internet_audio_players", "2007_establishments_in_Germany", "Virtual_communities", "Internet_properties_established_in_2007", "Music_websites", and their "sameAs". In this case, as "Music" belongs to "Intent" and "Interest" categories of initial user profile ontology, "hasIntent" and "hasInterest" relationships are added.

Something analogous happens to domain *www.telecinco.es* linked to "Television" category provided by OpenDNS. User profile ontology is populated with new concepts such as "Companies_based_in_the_Community_of_Madrid", "Television_channels_and_stations_established_in_1990", "Television_stations_in_Spain", "Grupo_Vocento", "Telecinco", "Spanish-language_television_stations", and their "sameAs" provided by searches in DBpedia knowledge base. Additionally, as "Television" belongs to "Interest" category of base user profile ontology, new "hasInterest" relationships populated the user profile ontology.

However, *www.tablondeanuncios.com, Unfollowers.me* and *t.co* are URLs that are neither in DBpedia nor OpenDNS. It should be mentioned that most of these domains are already categorized into OpenDNS but are "awaiting votes" by Domain Tagging community, waiting to be confirmed in an "approved" category. At this time, our user profiling algorithm takes into account only the "approved" categories by the community. It provides a more accurate categorization of URLs.

As a result, all relationships generated are stored and can be queried in Virtuoso database, where billions of relationships can be asserted without performance problems. Experimental results based on user's tweets confirm that the proposed method improves the automatic acquisition of interests and intentions.

VI. Conclusions and Future Work

Information on the Internet is growing exponentially and in recent years, social networks are being widely used. How to provide useful personalized services to adapt the interests and intentions of users become a critical issue and an important business strategy. In this paper, an approach based on the generation of user profiles is proposed in order to provide advertisements and services that really interests.

According to the results obtained, we may face new tests and analysis of results with optimism. Our approach allows obtaining automatically the interests and intentions of users through the URLs they share. They also show that the support of additional knowledge bases such as OpenDNS and DBpedia plays an important role and has a significant positive effect on user profiling. In addition, social networks and big data require store, analyze and process large amount of data in distributed databases that provide high availability and scalability.

The aim of the future work is to further investigate in massive data processing and clustering, allowing the implementation of scalable algorithms. At present, we are working on analyzing the performance and scalability of the method presented, exploring the possibilities of massive data and user processing. Moreover, we are working on including the time dimension to capture the changes of user profiles, the categorization and opinion analysis of the Twitter text and the incorporation of new knowledge repositories such as DMOZ[9] or SUMO[10] in order to improve and enrich the user profiling.

Acknowledgment

This work has been partly sponsored by the Mechatronics and System Group (SISTRONIC) of the Aragon Institute of Tecnhnology and the projects "*NOVARED: Sistema para Distribuir e Incrementar el Valor Creado en Internet"* and "*QuEEN*: *Quality of Experience Estimators in Network"*.

[9] Open Directory Project (http://www.dmoz.org/)

[10] Suggested Upper Merged Ontology, SUMO (http://www.ontologyportal.org/)

REFERENCES

[1] Malhotra, A., Totti, L., Meira Jr., W., Kumaraguru, P. & Almeida, V. "Studying user footprints in different online social networks". 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp.1065–1070 (2012).

[2] Tao, X., Li, Y., Lau, R. Y. K. & Geva, S. "Ontology-based specific and exhaustive user profiles for constraint information fusion for multi-agents". 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 264–271 (2010).

[3] Aïmeur, E., Brassard, G. & Molins, P. "Reconstructing profiles from information disseminated on the Internet". 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, pp. 875–883 (2012).

[4] Pennacchiotti, M. & Popescu, A. "Democrats, republicans and starbucks afficionados: user classification in twitter". Proceedings of the 17th ACM SIGKDD, pp. 430–438 (2011).

[5] Pennacchiotti, M. & Popescu, A. "A machine learning approach to twitter user classification". Proceedings of the Fifth International AAAI Conference on Weblogs and SocialMedia, pp. 281–288 (2011).

[6] Dey, L. & Gaonkar, B. "Discovering regular and consistent behavioral patterns in topical tweeting". 21st Internatinal Conference on Pattern Recognition (ICPR 2012), pp. 3464–3467 (2012).

[7] Wagner, C., Liao, V., Pirolli, P., Nelson, L. & Strohmaier, M. "It's not in their tweets: modeling topical expertise of Twitter users". 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, pp. 91–100 (2012).

[8] Quercia, D., Kosinski, M., Stillwell, D. & Crowcroft, J. "Our Twitter profiles, our selves: predicting personality with Twitter". 2011 IEEE International Conference on Privacy, Security, Risk and Trust and 2011 International Conference on Social Computing, pp. 180-185 (2011).

[9] Siehndel, P. & Kawase, R. "TwikiMe! - User profiles that make sense". 11th International Semantic Web Conference: ISWC2012 (2012).

[10] Tao, K., Abel, F., Gao, Q. & Houben, G. TUMS: "Twitter-based user modeling service". The Semantic Web (ESWC 2011), pp. 269–283 (2012).

[11] Sakaki, T., Okazaki, M. & Matsuo, Y. "Earthquake shakes Twitter users: real-time event detection by social sensors". 19th International World Wide Web Conference Committee (IW3C2), pp. 851–860 (2010).

[12] Lee, B. & Hwang, B.-Y. "A Study of the correlation between the spatial attributes on Twitter". 2012 IEEE 28th International Conference on Data Engineering Workshops, pp. 337–340 (2012).

[13] Li, R., Wang, S., Deng, H., Wang, R. & Chen-Chuan C., K. "Towards social user profiling: Unified and discriminative influence model for inferring home locations". ACM International Conference on Knowledge Discovery and Data Mining, pp. 1023–1031 (2012).

[14] Lauschke, C. & Ntoutsi, E. "Monitoring user evolution in Twitter". 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 972–977 (2012).

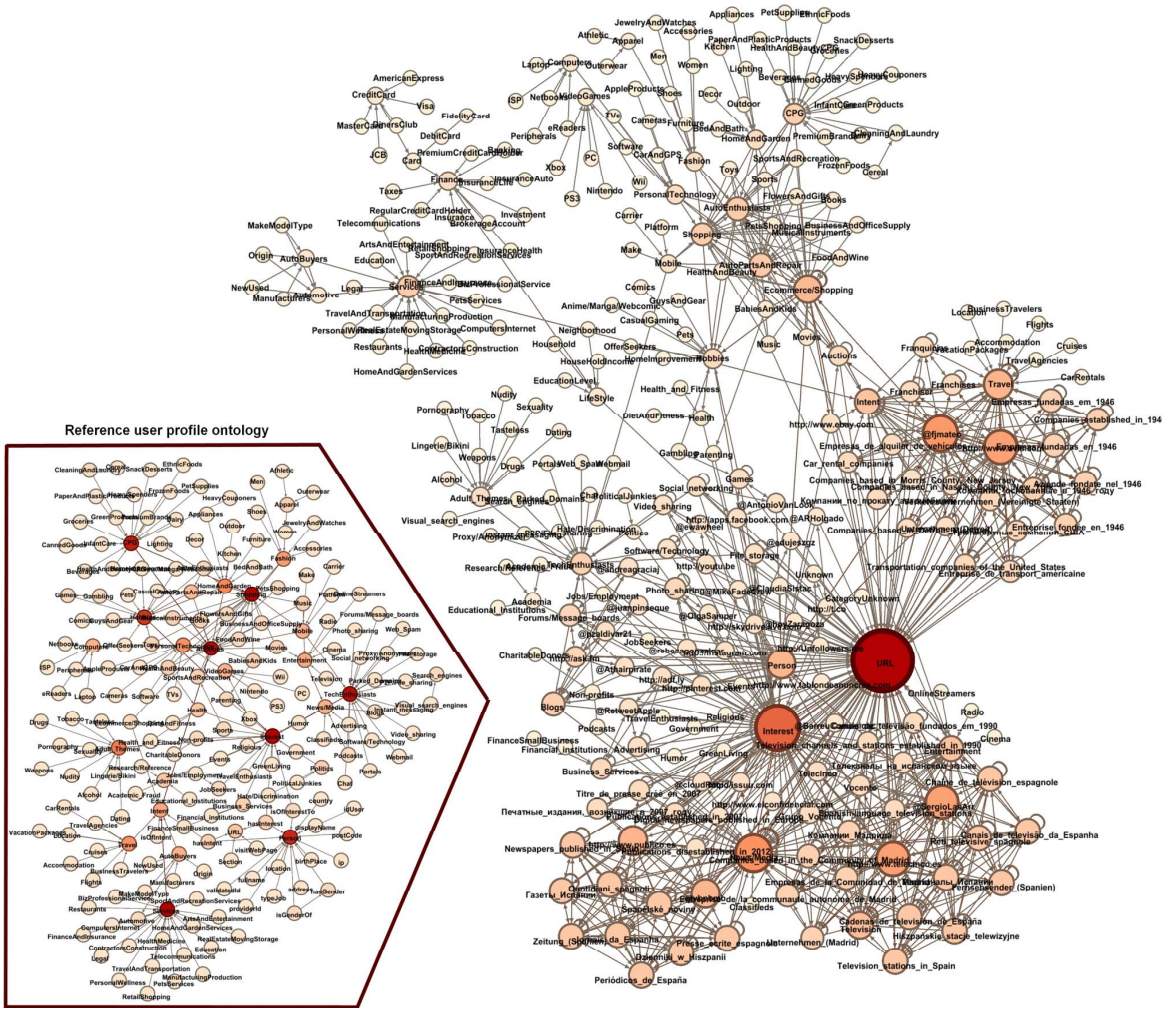

Fig. 5. Topology of reference ontology and topology of inferred model generated