Leveraging Social Network Analysis with Topic Models and the Semantic Web

Sebastián A. Ríos
Industrial Engineering Department
University of Chile
srios@dii.uchile.cl

Felipe Aguilera
Computer Science Department
University of Chile
faguiler@dcc.uchile.cl

Francisco Bustos
Computer Science Department
University of Chile
fbustos@dcc.uchile.cl

Tope Omitola
Intelligence, Agents, Multimedia (IAM) Group
School of Electronics and Computer Science,
University of Southampton, UK
tobo@ecs.soton.ac.uk

Nigel Shadbolt
Intelligence, Agents, Multimedia (IAM) Group
School of Electronics and Computer Science,
University of Southampton, UK
nrs@ecs.soton.ac.uk

***Abstract*—Social Network Analysis (SNA) and Web Mining (WM) techniques are being applied to study the structures of social networks in order to manage their dynamics and predict their evolution. This paper describes how we used Semantically-Interlinked Online Communities (SIOC) ontology to represent the (latent) semantic relationships between the members of a large community forum (about 2,500), Plexilandia. We extended SIOC, taking advantage of topic-based text mining and developed data mining algorithms that used our SIOC extensions to provide a better understanding of the social dynamics of the members of the Plexilandia community. This new understanding helped us to detect and discover the key members of Plexilandia successfuly.**

***Keywords*-Semantic Web; Online Social Networks; Text Mining; Topic Models**

I. Introduction

The web is now a major communication platform used by all sections of society, connecting people, organizations, and knowledge through their objects-of-interest, forming "object-centred networks" [1]. As the web gets more embedded and pervasive, these object-centred networks become bigger and more disparate, leading to a surfeit of interactions and data but very little knowledge. These object-centred networks are inherently social networks, and the people within these networks need to understand what information is accessible in order to know where the gaps are, whom they need to ask for more information, etc.

This work focuses on how we have extended the SIOC ontology to store important conceptual information from online community members. We used this extension to enhance the structure of the social relations graphs. This allowed us to develop new data mining techniques and to leverage the power of Social Network Analysis (SNA). We used a topic model driven text mining process to obtain the conceptual information of the social network. We obtained social network graphs from our RDF N3 instance data using SPARQL[1]. We tested our approach on a large community called Plexilandia[2]. We computed key members and sub communities based on the new information gathered from the social networks. This way, we were able to discover new hidden relationships among social network members.

II. Related Work

A vital task for administrators when enhancing the organization of a virtual community's contents and links structure, is to analyze data generated by that community. However, massive amounts of data may need to be analyzed in short periods of time (an hour or in a couple of hours). To solve this issue, SNA is being used to perform automated analysis to gather valuable information of community structure (experts, key members, sub groups, passive members, etc.) based on relationships between community members [2], [3]. Others prefer to use data mining, especially web mining (WM), where the structure of social interactions is lost but it is possible to find interesting patterns of texts in members posts or navigation patterns [4], [5], [6].

A. Efforts to add Semantics into Social Network Analysis

SNA helps to understand relationships in a given community from analyzing its graph representation. Users are seen as nodes and relations among users are seen as arcs. This way, several techniques have been proposed to extract key members [7], classify users according their relevance within the community [8], discovering and describing resulting subcommunities [9], etc. However, all these approaches leave aside the meaning, (i.e. semantics) of relationships among users.

Pathak et al. [10] proposed a community based topic-Model integrated social network analysis technique (Community–Author–Recipient–Topic model or CART) to extract communities from an emails' corpus based on the

[1]http://www.w3.org/TR/rdf-sparql-query/
[2]http://www.plexilandia.cl/foro/

topics covered by different members of the overall network. Similarly [4] developed an LDA–driven process to extract meaningful topics for members of a virtual community of practice in order to discover for example hidden key-members.

III. The Social Web

A. Adding Semantics to the Web

The Web has always been a set of resources (such as web pages, files, etc) connected to each other by hypertext links which are untyped. By untyped we mean that the relationships between the linked resources are not easily discernible and are not necessarily machine-interpretable. Adding meaning (i.e. semantics) to these resources and their linkages will make them more machine interpretable thereby making it easier to find relevant information from related social structures. Ontologies and ontology languages are the key concepts used in this regard to provide additional meaning to these resources and their links. RDF[3] and OWL [4] are the main ontology languages used to represent meaning on the Semantic Web.

B. Modeling Online Social Networks

Semantic Web frameworks provide a graph model (RDF), a query language (SPARQL[5]), and type and definition systems (RDFS[6] and OWL) to represent knowledge on the Web. These frameworks provide a better way to represent and model social networks in much richer structures than plain (SNA) graphs.

1) Ontological representation of individuals: Several ontologies, e.g. the Friend-of-a-Friend (FOAF) ontology, can be used to represent individuals. FOAF describes a widely-used vocabulary for describing people, the relationships between them, and their activities (i.e. what they create and do).

2) Ontological representation of online social communities: The SIOC ontology[7] describes a vocabulary for describing interlinkages between related online community content from platforms such as blogs, message boards, and other social websites. In combination with the FOAF vocabulary for describing people and their friends, and the SKOS[8] model for organizing knowledge, SIOC lets people link discussion posts and content items to other related items and discussions, to people, and topics.

C. Semantic Social Network Analysis for Data Mining

One of the uses of SNA is community discovery. A community is simply a group of entities that shares a common interest or is involved in an activity or event. We believe that a community is comprised of the people, in that community, the goals and purposes, of that community, and the policies of members' interactions, in that community. FOAF provides the concepts for representing people and their activities, but does not have the concepts that can be used to represent a community's purposes and policies. SIOC, on the other hand, provides the vocabulary for representing an online community but is missing the vocabulary for representing community's purposes, its policies, and the participation levels, e.g. the expertise levels of community's members. These make it difficult to perform efficient sub-community detection/discovery using these ontologies.

In view of this, we provide an extension to SIOC, called "**sioc-sna-dm**", that makes it possible to represent a community's purposes, its policies, and its goals. sioc-sna-dm[9] comprises the following major modules and classes:

1) People: This is made up of (a) Social Entity representing a social entity, such as an actor or a group of actors, (b) Group, which represents a group or set of actors, (c) Person, which represents a person or an actor, (d) Role, this represents the role performed by the social entity, and (e) Interaction,
2) Policies composed of (a) Permissions representing the permissions accorded a particular role played by a social entity, (b) Personal Moderation, (c) Interaction Moderation, and (d) Context Moderation,
3) Purposes, which consists of (a) Tags given to a particular post by a social entity, and can also be automatically generated by LDA, (b) Container, and (c) Context.

D. Generating RDF instances from Plexilandia database

We used the sioc-sna-dm ontology as a template to generate the RDF instances from the Plexilandia database (DB). We adopted the N3 format of RDF due to its terseness when compared to other RDF formats such as RDF/XML. The Plexilandia DB was a MySQL DB that had data of users, users' details, their posts, as well as records of their interactions. Although we were aware of R2RML (the RDB to RDF Mapping Language)[10], the specification was still very new at the time of our work and very few implementation of it existed. We developed our own RDB-to-RDF conversion program, written in Java, that used JDBC to connect to Plexilandia DB, getting the MySQL data from the database and converted these to RDF/N3. A snippet of the RDF for a social interaction between two members is shown below (lda represents the latent Dirichlet allocation value for this interaction).

```
<http://www.plexilandia.cl/foro/
   socialinteraction/66/100245>
a  def:SocialInteraction ;
```

[3] http://www.w3.org/TR/2004/REC-rdf-primer-20040210/
[4] http://www.w3.org/TR/owl-ref/
[5] http://www.w3.org/TR/rdf-sparql-query/
[6] http://www.w3.org/TR/rdf-schema/
[7] http://rdfs.org/sioc/spec/
[8] http://www.w3.org/2004/02/skos/
[9] sioc-sna-dm's ontology is at http://www.enakting.org/provenance/siocSnaDm/.
[10] http://www.w3.org/TR/2010/WD-r2rml-20101028/

```
def:lda "0.236321"^^xsd:string ;
def:original_post_user_identifier
"100245"^^xsd:string ;
 def:reply_post_user_identifier
 "101262"^^xsd:string .
```

E. Topic-based Text Mining using LDA

A topic model can be considered as a probabilistic model that relates documents and words through variables which represent the main topics inferred from the text itself. In this context, a document can be considered as a mixture of topics, represented by probability distributions which can generate the words in a document given these topics.

A main topic model is the Latent Dirichlet Allocation (LDA) [11]. The key idea of LDA, is that every topic is modeled as a probability distribution over the set of words represented by the vocabulary and every document as a probability distribution over a set of topics. These distributions are sampled from multinomial Dirichlet distributions.

IV. Application to Plexilandia Community

We performed several experiments on the web site of plexilandia.cl, a Virtual Community of Practice (VCoP). Our evaluation was based on interviews with the administrators plus a preliminary study of community activity. This was presented in [4], [5]. However, in this paper we use our semantic representation **(sioc-sna-dm)** to reconstruct the same experiments to probe that the model supports social network analysis and also data mining on a VCoP which is a specialized (more restricted) social network.

A. Plexilandia

Plexilandia, a VCoP of over 9 years old with over $2,500$ members, meet to build music effects, amplifiers, and audio equipment. Although, they have a basic community information web page, most of their members' interactions are produced on the discussion forum. Since its inception, the site's administration staff has increased from 1 to 5. Today, the administration tasks are typically: Items Classification, User Moderation, and User Participation.

The vision of administrators and experts about the community is based mostly on experience and time participating in the community. They also have some basic and global measures. For example, total number of posts, connected members, etc. However, they do not have: members browsing behavior information, members publications' quality and how they contribute to the purpose of the community.

B. Extracting Filtered Networks From SIOC-SNA-DM

Using our semantic representation, we are able to extract a traditional graph representation of Plexilandia. We count the number of messages that a user sends to another user in the traditional way. We call this "counting method", see Fig. 1(a). In this image we can observe a graph representation without any filter method. Afterwards, using our extended **sioc-sna-dm** ontology model, we are able to extract a topic filtered network, see Fig. 1(b), since this information is also stored in our model.

The benefits of the above is, for instance, that a filtered network has a density of 0.004 compared to the counting network which has a density of 0.019. Therefore, we obtained a density reduction of about 80% and we kept the quality of traditional methods results (as shown in [4], [5]). Besides, since we also store intrinsic topics information, we can also enhance other algorithms outputs, as we will show in the next sections. The SPARQL snippets below show how we got the data for the graph representation:

```
SELECT DISTINCT ?originalPostUser
?replyPostUser  ?ldaValue where {?s a
<http://purl.org/net/sioc-sna-dm-uchile/
     def/SocialInteraction> .
?s <http://purl.org/net/sioc-sna-dm-uchile/
 def/original_post_user_identifier>
?originalPostUser . ?s
<http://purl.org/net/sioc-sna-dm-uchile/
 def/reply_post_user_identifier>
?replyPostUser .
?s <http://purl.org/net/sioc-sna-dm-uchile/
def/lda> ?ldaValue . };
```

C. Key-members Discovery

In a VCoP, the existence of members that are experts in a field is of tremendous importance, since they are the ones who share their experience and knowledge. There are several algorithms to detect key-members of a community, but the most commonly used are HITS and PageRank. Based on graph representations that we obtained before, we chose gephi[11] to perform SNA on these graphs. We used HITS hub to rank key-members (table I). We showed we can regenerate the results from our previous works.

Table I
Top 10 Key-members with both approches

NOT FILTERED		FILTERED	
USERS	HUB	USERS	HUB
sfehland	0.02865497	NIOpren	0.013422819
Belzeboo	0.025730994	strat_cl	0.013422819
robotekmania	0.022807017	FeDomic	0.013422819
Cristian74	0.022222223	Daido	0.013422819
jotamachuca	0.021637427	mcruli	0.013422819
luchin542	0.021637427	douglaz	0.013422819
vaitriani	0.021052632	Belzeboo	0.013422819
Pulentboy	0.018128656	felipe_blues	0.013422819
jsanta	0.018128656	Oragall	0.013422819

D. Sub-community Discovery

Other important feature is the discovery of new subgroups based on the graph representation of the VCoP. However, since we store semantic information regarding the main topic of every post in the community using LDA, we are able to automatically create sub groups of members based on their topic of interests. Moreover, the modularity index calculated on the traditional (not filtered) graph is 0.21

[11] http://gephi.org/

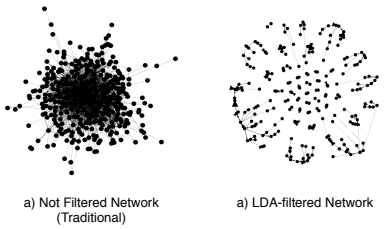

Figure 1. Creator Oriented Network for 2009 (Force Atlas Visualization)

which led to 31 communities in our data. However, on the LDA-filtered network, modularity is 0.78 which indicates that there are around 78 sub communities[12]. This way, we observe that filtering the network allows us to discover new hidden sub communities which were not possible before using traditional representation. Since sioc-sna-dm considers intrinsic and explicit topics of interactions, it thus enables these advanced SNA techniques and Text Mining techniques to be applied.

V. Conclusion

In this paper we have extended the SIOC (Semantically-Interlinked Online Communities) ontology. A main contribution of this work is our proposed model based on three main characteristics applicable for all communities, and these are: **People, Policies,** and **Purposes**. Moreover, the extended model can store information regarding the inner semantics of the contents of the interactions of the members of a community. In particular, we have provided an example of how to populate this ontology using an LDA based text mining and a concept-driven text mining approach. However, the proposed model is generic, and any topic model can be used. This way, we can leverage the techniques that can be applied to perform Social Network Analysis (SNA) and also Data Mining (DM) on a social network. We performed a real application to a community called plexilandia that started on 2002 and now has more than 2500 members.

Acknowledgments

We thank grants ICM: P-05-004-F, CONICYT: FBO16 and project code: 11090188, entitled "Semantic Web Mining Techniques to Study Enhancements of Virtual Communities" (www.snagroup.cl). We thank the EnAKTing project, funded by EPSRC project number EP/G008493/1.

References

[1] S. Kinsella, A. Passant, J. G. Breslin, S. Decker, and A. Jaokar, "The future of social web sites: Sharing data and trusted applications with semantics," *Advances in Computers*, vol. 76, pp. 121–175, 2009.

[2] J. Qiu, Z. Lin, C. Tang, and S. Qiao, "Discovering organizational structure in dynamic social network," *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pp. 932–937, 2009.

[3] L. Tang, H. Liu, J. Zhang, and Z. Nazeri, "Community evolution in dynamic multi-mode networks," *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 677–685, 2008.

[4] H. Alvarez, S. A. Ríos, F. Aguilera, E. Merlo, and L. Guerrero, "Enhancing social network analysis with a concept based text mining approach to discover key members on a virtual community of practice," *Lecture notes in computer science*, vol. 6277, pp. 591–600, 2010.

[5] G. L'Huillier, H. Alvarez, S. A. Ríos, and F. Aguilera, "Topic-based social network analysis for virtual communities of interests in the dark web," *SIGKDD Explorations Newsletter*, vol. 12, no. 2, pp. 66–73, 2010.

[6] S. A. Ríos, F. Aguilera, and L. Guerrero, "Virtual communities of practice's purpose evolution analysis using a concept-based mining approach," *Knowledge-Based Intelligent Information and Engineering Systems - Part II; Lecture Notes in Computer Science*, vol. 5712, pp. 480–489, 2009.

[7] R. D. Nolker and L. Zhou, "Social computing and weighting to identify member roles in online communities," *Web Intelligence, IEEE / WIC / ACM International Conference on*, vol. 0, pp. 87–93, 2005.

[8] U. Pfeil and P. Zaphiris, "Investigating social network patterns within an empathic online community for older people," *Computers in Human Behavior*, vol. 25, no. 5, pp. 1139–1155, 2009.

[9] H. Kwak, Y. Choi, Y. H. Eom, H. Jeong, and S. Moon, "Mining communities in networks: a solution for consistency and its evaluation," in *Proceedings of the 9th ACM Internet Measurement Conference (IMC)*. New York, NY, USA: ACM, Nov. 2009, pp. 301–314.

[10] N. Pathak, C. DeLong, A. Banerjee, and K. Erickson, "Social topic models for community extraction," *The 2nd SNA-KDD Workshop*, vol. 8, 2008.

[11] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, Jan 2003.

[12]Modularity was calculated using gephi