Neurocomputing 326–327 (2019) 71–81

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Semantically enhanced network analysis for influencer identification in online social networks

Sebastián A. Ríos[a], Felipe Aguilera[b], J. David Nuñez-Gonzalez[c], Manuel Graña[c,*]

[a] Department of Industrial Engineering, Universidad de Chile, Republica 701, P.O. Box: 8370439 Santiago, Chile
[b] Department of Computer Science, University of Chile, Blanco Encalada 2120, Santiago P.O. Box: 8370459, Chile
[c] Computational Intelligence Group, University of the Basque Country, Paseo Manuel Lardizabal 1, San Sebastian 20018, Spain

ARTICLE INFO

Article history:
Received 30 August 2016
Revised 27 December 2016
Accepted 29 January 2017
Available online 20 September 2017

Keywords:
Online Social Networks
Social network analysis
Latent topic analysis
Semantic modelling
Fuzzy concept analysis
Influencer detection

ABSTRACT

Influencers in a social network are members that have greater effect in the online social network (OSN) than the average member. In the specific social networks known as communities of practice, where the focus is an specific area of knowledge, influencers are key for the healthy working of the OSN. Approaches to influencer detection using graph analysis of the network can be mislead by the activity of users that are not contributing to the OSN purpose, bogus generators of documents with no relevant information. We propose the use of semantic analysis to filter out such kind of interactions, achieving a simplified graph representation that preserves the main features of the OSN, allowing the detection of true influencers. Such simplification reduces computational costs and removes bogus influencers. We demonstrate the approach applying fuzzy concept analysis (FCA) and latent Dirichlet analysis (LDA) to compute document similarity measures that allow to filter out irrelevant interactions. Experimental results on a community of practice are reported.

1. Introduction

The advances in communications and web services has fostered the creation of Online Social Networks (OSN) which allow people to carry out remote social interactions regardless of geographical and even cultural distances, leading to the emergence of new social institutions [32,34,36,37] which, although, are based on existing ones, possesses specific characteristics [10]. For instance, many face-to-face social rituals [33] do not exist or have been adapted to the virtual world [9]. Specifically, the communities of practice [24] that share knowledge about a specific activity have benefitted from the new social media, Knowledge sharing [12] is performed through blogging or forum services, where users post their questions, responses, and proposals for cooperative project development.

In order to monitor the healthy operation of the OSN, system administrators resort to Social Network Analysis (SNA) techniques, in order to gather valuable information about social network structure (experts or influencers [7], sub groups, passive members, etc.) based on relationships between social network members formalised as a graph representation of the OSN. With the huge increase in size of the OSN, SNA tasks pose strong problems for real time processing, which may be required in some extreme circumstances. Therefore, simplification approaches to the OSN graph representation that preserve fundamental properties highly desirable. For specific forms of OSN, such as the communities of practice, the semantic analysis of the documents posted and exchanged by the users offer a promising avenue for innovation [1,14,18,26]. Semantic analysis may allow to remove irrelevant user interactions, either because they are not related or they do not offer additional information, so that the OSN graph representation is effectively simplified.

We have implemented two different kinds of semantic analysis: fuzzy concept analysis (FCA) and latent Dirichlet allocation (LDA) [2]. The ability to simplify the OSN graph of both methods tested proved over three different strategies to extract the OSN graph representation from the documents posted in its forum. We examine the graph density reduction achieved by each approach. The OSN analysis consists in the discovery of influences (aka key-members) applying the well known authority discovery algorithm (ADA) [11]. Carrying the ADA on the unfiltered and the semantic filtered graphs we assess that the later do not degrade the key-member discovery.

This paper is organized as follows. Section 2 discusses the related work in the literature. Section 3 presents the mathematical formalisation of the graph construction, and the semantic

* Corresponding author.
E-mail addresses: srios@dii.uchile.cl (S.A. Ríos), faguiler@dcc.uchile.cl (F. Aguilera), manuel.grana@ehu.eus (M. Graña).

https://doi.org/10.1016/j.neucom.2017.01.123
0925-2312/

analyses carried out. Section 4 describes the experimental setup on a real world OSN, giving the computational results. Finally, Section 5 presents our conclusions together with some ides for future work.

2. Related work

There are different issues to research around OSN's. For example, communities detection [6], moderation administration [8], and, relating to present work, key-members analysis. For OSNs it is very important to help generate, store, and preserve knowledge resulting from member interaction. The success of a OSN depends on appropriate administration tools [25], and the activity of influencers (aka leaders [3], core members [25], or key-members). Likewise, the goal of every OSN members is to acquire specific knowledge from the social network.

There are many definitions of influencer, e.g. the most participative members [25], the members who answer other members questions [15], or the member who encourage others members to participate [3] proposing projects or dissuasion topics. However, none of these definitions take into consideration the content or the meaning of these interactions carried out by means of documents (posts/comments, reply, questions asked, reports created, etc.) that can be analysed. An interaction is taking into account just because user A replies a post from user B. Whether if A answer to the question is satisfactory or not, it is not taken into account. Those approaches just consider equally each post in a sequence, even if the post topic is irrelevant to the thread which contains it.

Influencer detection is a very important administrator task, because these members keep the OSN alive. They share their experiences, knowledge, create tutorials, develop videos on a subject to help other non-experts members, etc. Often, administrators or social network owners, may pay these experts to develop contents for the OSN, since t these contents will increase the social network impact, fostering high interaction between members, and helping to capture new members. In small OSNs, where the quantity of both members and posts can be checked manually, many times administrators or owners know almost all members and their participation. Therefore, they all know who-is-who in the social network. However, as the OSN becomes bigger, where thousands of members publish thousands of posts daily, without specific computational tools, tracking the users activity becomes an unmanageable task.

Concept-Based Text Mining has been applied [26] to track the evolution of the purpose of a Virtual Community as a measure of community goal accomplishment, obtaining a set of scores that show how relevant, relative to its stated purpose, is the content generated by the community. This approach could be adapted into a *"users goal accomplishment score"* for each post, presenting a measure of the interaction between members which consider the content of theirs replies.

Using a graph representation of the social network, based on the replies between members [29] analyse the role of the middleman members, aka brokers. These members are the link between askers and repliers. The weight of an arc is a measure of member's interaction. They found the evolution through time of the brokers and how valuable are they for the social network. However, they do not consider the semantics of the information exchanged.

Related to the analysis of Virtual Communities [5] performs a statistical work to establish why a member keeps his knowledge-sharing intention through time. They made a survey in which three research streams related to the member's intentions are measured. They not only measure the trust between members, they also prove that trust between members and network administration has influence in the community too. This approach it is very important, because a healthy community depends of the knowledge-sharing intention of the members, specially of the key-members. Furthermore [4] does a similar work as the latter, with the aim to explain why members give or receive knowledge to/from other community members. Interpersonal trust and knowledge sharing self-efficacy are viewed as the most relevant contributions in a community. These works emphasize that trust between members is important to have a healthy community. Therefore, key-members play a relevant work for social knowledge generation, motivating other members to participate and frequently bringing up new key-members.

On the other hand [30] follows a marketing analysis approach to determinate the influence that Social Networks have in customer consuming decisions. They identified six categories of members according his interest and participation in the social network, remarking the core members who are the most frequent visitors and the ones who spend more time sharing his knowledge and participating in different threads.

SNA [31] helps to understand relationships in a given community analysing 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 influencers or experts members [20], classify users according his relevance within the social network [22,38], discovering and describing resulting subcommunities [13], among other applications. One limitation pure SNA, such as the authoritative node detection (ADA) [11], based on measured interaction only is that ignoring the content created by users could result in detecting as experts users that behave like flooders, trolls or spammers, and worst, disregarding the real social network experts. McCallum et al. [16,17] described how to determine roles and topics in a text-based social networks by building Author-Recipient-Topic (ART) and Role-Author-Recipient-Topic (RART) models. Furthermore, in Pathak et al. [21], a community based topic-Model integrated social network analysis technique (Community-Author-Recipient-Topic model or CART) is proposed to extract communities from a emails corpus based on the topics covered by different members of the overall network. These approaches novelty is the use of data mining on text from the social network to perform SNA to study Roles or Sub-groups extraction. This is to our knowledge one of the best automated topics extraction approach to perform key-members detection and later used to role discovery. However, it was applied only on e-mail networks where there isn't a clear purpose or goals to accomplish.

3. Graph construction and semantic analysis

Our main research question in this paper is how to enhance the discovery of influencers in the social network structure using semantic information. In an OSN supporting a community of practice, the most important interaction device is the forum tool, where users post queries, projects, responses and any kind of contribution (hopefully) related to the social network main purpose. The construction of the graph representation of the OSN is based on the forum interactions, so that nodes correspond to active users and arcs between nodes correspond to the interactions in the form of replies to posted documents. Next subsection discusses alternative definitions of the baseline graph representation, taking into consideration post responses in different ways. The semantic analysis of the posted documents is used to filter out responses that or irrelevant or noisy. Hence we define normalized semantic similarities between documents which are used to detect irrelevant documents as outliers with very small similitude to the original post, which are removed from the graph representation. Semantic analysis is based on the probabilistic encoding of the documents on the basis of an accepted vocabulary, which is explained before dealing with fuzzy concept modelling and latent dirichlet analysis (LDA).

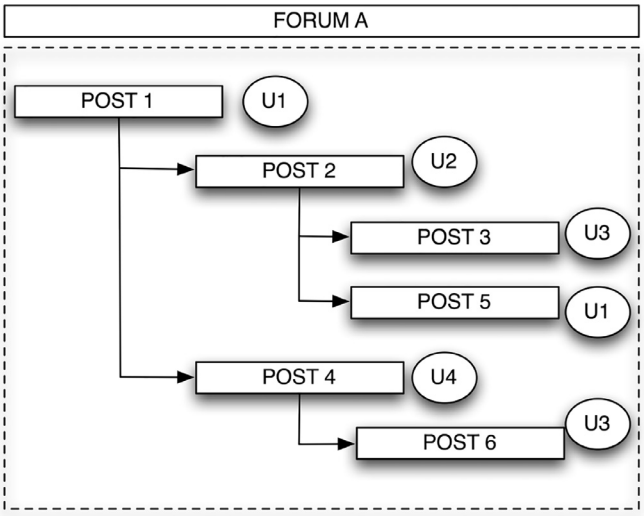

Fig. 1. A typical social network forum dialog structure. Circles represent the users who posted the documents represent by rectangular boxes. Arrows represent the response relation between posts.

3.1. Identification of key-members

The identification of the social network core constituted by the key-members on the OSN graph representation is carried out applying a classical algorithm, i.e. the ADA algorithm [11]. The ADA was formulated to detect authoritative web pages to present the core contents of a topic in response to a query of a user. The algorithm is based on a method for the localisation of dense bipartite communities in the graph structure. Looking for authoritative sources, the method finds two mutually reinforcing sets of nodes: authorities and hubs. The nodes in the second set point to qualitatively better authority nodes, while being unrelated themselves. Authorities are large in-degree nodes which are pointed to by hub nodes. These are the bipartite communities searched for by the algorithm. The algorithm proceeds by a simple iteration of weights assignments which can be proven to converge to an equilibrium state. Each node has both an authority and hub weight. Hub weights are updated by the sum of the authority weights of nodes at the other end of outgoing arcs. Conversely, authority weights are updated by the sum of the hub weights of nodes at the other end of incoming arcs.

3.2. Graph representations of the social network

The graph representation of the social network is built taking into consideration members interaction. In general, members activity is measured according to their participation on the social network forum. A member participation is detected when he/she posts a document in the social network. Therefore, graph nodes will be the OSN members. Because the overall activity of the OSN is measured according to members participation, Graph arcs will represent interaction between them. How to link the members and how to measure their interactions to built up the graph's network is our main concern in the following. Fig. 1 gives a schematic representation of a dialogs that may take place in the forum service of an OSN. Each circle is a user posting a document which may be a response to previous posts.

Three graph representation of OSN may be built from the dialog backlog according to the following edge creation strategies:

1. **Creator**: When a member create a thread of posted documents, every reply will be related to him/her. Therefore, the arcs are only defined between the user generating a reply and the user responsible for the original post in the thread. This network representation is the less dense of all the possible graph representations.
2. **Last Reply**: Every reply of a thread will be considered as a response to the last post only. Therefore, the arcs are only defined between the user generating a reply and the user responsible for the last post in the thread. This graph representation has a middle density.
3. **All Previous**: Every reply of a thread will be a response to all posts which are already in a specific thread. Therefore, the arcs are defined between the all user generating a reply and the users responsible for the all previous post in the thread. This graph representation is the densest network.

Fig. 2 illustrates these three approaches to build the graph representations for the user dialog shown in Fig. 1. The arcs weight is a counter of how many times a member replies to another one. We carry out concept and topic based message analysis in order to detect and analyse the reply of members according to the social network purpose (for all graph representations discussed above), and to filter out noisy posts. The following sections deal with the semantic filtering of the arcs in order to achieve less dense graphs which have lower computational requirements to produce comparable results in the identification of key-members. We focus on the networks built by the 'creator' strategy in this paper.

3.2.1. Vocabulary based document encoding

In the following we denote by $\mathcal{V}$ the vocabulary given by a collection of words $w \in \{1, \ldots, |\mathcal{V}|\}$ where a word is defined as the basic discrete unit of data. A Web Post (aka document) is a sequence of S words denoted by $\mathbf{w} = (w^1, \ldots, w^S)$, where w^s represents the s^{th} word in the message. Finally, a corpus is defined by a collection of $\mathcal{P}$ documents denoted by $\mathcal{C} = (\mathbf{w}_1, \ldots, \mathbf{w}_{|\mathcal{P}|})$. In matrix notation, the corpus of documents is encoded as $\texttt{TF-IDF} = \left[(m_{ij}), i \in \{1, \ldots, |\mathcal{V}|\}; j \in \{1, \ldots, |\mathcal{P}|\}\right]$, where m_{ij} is the weight of the i^{th} word in the j^{th} document. We define the m_{ij} weights as a *term frequency times inverse document frequency, tf-idf* [28], formally defined by:

$$m_{ij} = \frac{n_{ij}}{\sum_{k=1}^{|\mathcal{V}|} n_{kj}} \times log\left(\frac{|\mathcal{C}|}{n_i}\right), \tag{1}$$

where n_{ij} is the frequency of the ith word in the jth document, and n_i is the number of documents containing word i. The *tf-idf* variable is a weighted representation of the importance of a given word in a document belonging to a corpus. The *term frequency* (TF) is the weight of the word in the document, while the *inverse document frequency* (IDF) measures whether the word is frequent or not in the document.

3.2.2. Fuzzy concept analysis

Fuzzy concept analysis (FCA) of posted documents is achieved making use of linguistic variables [27]. The linguistic variable (LV) values are words or sentences in natural language, thus they can not be treated by conventional numerical algorithms. Let u be a LV, so that the set of terms $T(u)$ cover its universe of discourse U, e.g. $T(temperature) = \{cold, warm, hot\}$ or $T(pressure) = \{high, ok, low\}$. A document can be represented as a fuzzy relation $[Concepts \times WP]$, which is a matrix where each row is a concept and every column is a web post. To obtain such matrix we can rewrite this relation in a more convenient manner as follows:

$$[Concepts \times WP] = [Concepts \times Terms] \otimes [Terms \times WP] \tag{2}$$

were "Terms" are the words that can be used to define a concept, and "WP" refer to any word inside a document. In Eq. (2) the

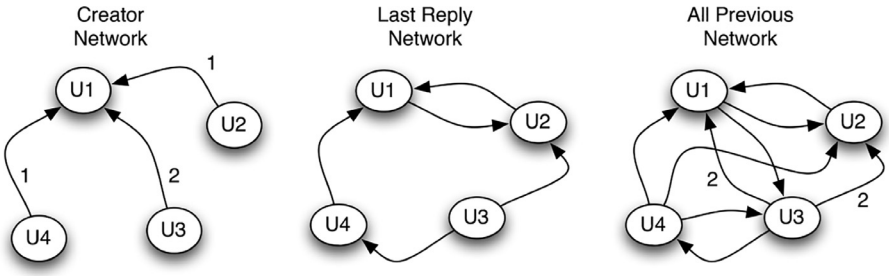

Fig. 2. Three different graph representation that may be created according to different strategies that represent the forum interaction illustrated in Fig. 1. In this paper we consider the CREATOR graph strategy.

operators " $\times$ " and "$\otimes$" denote the fuzzy relation and fuzzy composition, respectively. Let us denote by K the total number of concepts defined for the OSN site. Thus the matrix [*Concepts* $\times$ *WP*] is made of the the membership function values as shown in Eq. (3), where $\mu_{C\times WP} = \mu_{[C\times T]\otimes[T\times WP]}$ represents the membership function of the fuzzy composition in Eq. (2), and membership values belong to the interval [0, 1].

$$\mu_{C\times WP}(x,z) = \begin{pmatrix} \mu_{1,1} & \mu_{1,2} & \cdots & \mu_{1,|\mathcal{P}|} \\ \mu_{2,1} & \mu_{2,2} & \cdots & \mu_{2,|\mathcal{P}|} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{K,1} & \mu_{K,2} & \cdots & \mu_{K,|\mathcal{P}|} \end{pmatrix}. \tag{3}$$

There are several alternatives to define the fuzzy composition [19]. One important issue that must be considered is that even if some words are missing in a document, the degree of expression of a concept should not suffer alterations. This is achieved by the following compositional rule [19]:

$$\mu_{Q\circ Z} = \bigvee\{\mu_Q(x,r) \wedge \mu_Z(r,y)\}. \tag{4}$$

To understand Eq. (4), let us assume two fuzzy relations $Q(U, V)$ and $Z(V, W)$ sharing a common set V. The membership functions for Q and Z are given by $\mu_Q(x, r)$ and $\mu_z(r, y)$, respectively, with $x \in U$, $r \in V$, and $y \in W$. The operator $\bigvee$ denotes the limited sum defined by $\bigvee(x, y) = min(1, x + y)$, and $\wedge$ is the algebraic product $= \wedge(x, r) = (x \cdot r)$.

In order to apply the above equations, we begin identifying the relevant concepts that characterize visitors' alignment to the purposes of the OSN using experts' knowledge about the most relevant concepts to describe visitors' behaviour in the web site. Subsequently, with the help of a thesaurus and dictionaries we extract words defining the relevant concepts i.e. we express every concept as a list of words (assuming that a concept is a LV) looking for synonyms, quasi-synonyms, antonyms, etc. Afterwards, we need to define the membership values for the fuzzy relations [*Concepts* $\times$ *Terms*] and [*Terms* $\times$ *WP*]. We computed the relative frequency of words in a web page to represent the membership values of matrix [*Terms* $\times$ *WP*]. For the definition of [*Concepts* $\times$ *Terms*] values, we asked the expert to assign a concept representation degree to each word. To carry out this task we followed an indirect method. The expert compared two terms each time and gave a value between 0 and 1. For example, a synonym can receive a value near 1; a quasi-synonym, may receive a value near between 0.65 and 1; an antonym can be set to 0, etc. Finally, we obtained the fuzzy relation $\mu_{G\times P}(x, z)$ applying Eq. (4).

3.3. Topic extraction using generative models

A topic model is a probabilistic model that relates documents and words through latent variables which encode the main topics inferred from the text itself. In this context, a document is a mixture of topics, each topic modelled by a probability distribution which can be used to generate the words in a document. The topic modelling approach that we will use is the Latent Dirichlet Allocation (LDA) [2]. LDA assumes a Bayesian approach where latent document topics are inferred from probability distributions estimated over the training dataset. In the LDA generative model, every topic is modelled by a probability distribution over the set of words represented by the vocabulary ($w \in \mathcal{V}$), and every document by a probability distribution over the set of topics ($\mathcal{T}$). These distributions are multinomial Dirichlet distributions.

Given the smoothing parameters β and α, from a set of topics $\mathcal{T}$, and a topic mixture θ, LDA aims to estimate the joint probability distribution generating a document $\mathbf{w} = (w^1, \ldots, w^S)$ composed from a set of words S, given by the following expression:

$$p(\theta, z, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha)\prod_{s=1}^{S} p(z_s|\theta)p(w^s|z_s, \beta), \tag{5}$$

where $p(z_s|\theta)$ can be represented by the random variable θ_i, such that topic z_s is presented in document i ($z_s^i = 1$). By integrating Eq. (5) over the random variable θ and summing over topics $z \in \mathcal{T}$, a close expression of the marginal distribution of documents can be obtained as follows:

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{s=1}^{S}\sum_{z_s\in\mathcal{T}} p(z_s|\theta)p(w^s|z_s, \beta)\right)d\theta. \tag{6}$$

The final goal of LDA is to build a generative model with the distributions estimated from a given corpus of documents. There are several methods developed for inferring these probability distributions, such as variational expectation-maximization [2], a variational discrete approximation of Eq. (6) empirically used by [35], and a efficiently implementation of Gibbs sampling Markov chain Monte Carlo model [23].

3.4. FCA and LDA graph filtering

Previous work [26] proposes a method to evaluate the degree of accomplishment of community goals. Here we will use this approach to classify the members' posts according to OSNs' goals, which are defined by a set of terms composed of keywords or statements in natural language. We reduce the density of a graph representation by considering only posts which are relevant to the purpose of the social network, and conversation threads which are related one to each other. Therefore, if a member posed a question, we detect all posts that answered this question in the ensuing thread. Then, we compute the similarity of those answers to the original question they are replying to. The similarity measure is the cosine distance between the vector encoding of the documents. If the similarity is over a certain threshold θ, an interaction will be considered between them, irrelevant responses are dismissed. Afterwards, we count all post interactions between member A and member B for every thread in the forum. Formally, function *ARC* returns the weight of the interaction between user A and user B:

$$ARC(U_A, U_B) = \sum_{\tau=1}^{T} d_\tau(\mathbf{P}_A, \mathbf{P}_B), \quad (7)$$

where $\mathbf{P}_A$ and $\mathbf{P}_B$ are all the interactions via posted documents between members A and B, respectively, in a thread $\tau \in \{1 \dots T\}$, and T is the last thread of the forum. The distance d_τ is defined as follows:

$$d_\tau(\mathbf{P}_A, \mathbf{P}_B) = \sum_{i=1}^{|P_A|} \sum_{j=1}^{|P_B|} d_m\left(\mathbf{P}_{A\tau i}, \mathbf{P}_{B\tau j}\right) s.t \quad d_m\left(\mathbf{P}_{A\tau i}, \mathbf{P}_{B\tau j}\right) \geq \theta \quad \forall \mathbf{P}_{A\tau i}, \mathbf{P}_{B\tau j}, \quad (8)$$

$$d_m(\mathbf{P}_i, \mathbf{P}_j) = \frac{\sum_k w_{ik} w_{jk}}{\sqrt{\sum_k w_{ik}^2 \sum_k w_{jk}^2}}, \quad (9)$$

where $w_{\tau ik}$ is the score of topic/concept k in the posts of user i defined by the Semantic Weights Matrix (SWM), described in see Algorithm 1, on a thread τ. We compute $d_m(P_{\tau i}, P_{\tau j})$ only if P_j is a reply to P_i. After that, we add 1 to the weight of arc $a_{i,j}$ if $d_m(P_{\tau i}, P_{\tau j}) > \theta$. Finally, short posts or general messages like "yes", "good bye", "it helped me a lot" must be discarded. Therefore, in the computation of $ARC(.)$ function, we do not consider posts whose relevance is below ρ, where $\rho \in (0, 0.2]$.

Algorithm 1 Computing Semantic Weighted Matrix.

Require: $\mathcal{V}$ (Vocabulary)
Require: $\mathcal{P}$ (Posts)
Require: k (Number of Topics or Concepts)
Ensure: Semantic Weights Matrix SWM$[|\mathcal{P}|, k]$
1: TF-IDF$[|\mathcal{P}|, |\mathcal{V}|]$ (Eq. 1)
2: SM$[k\,\mathcal{V}] \leftarrow$ Build SM (semantic matrix) according to Topics (Sec. 3.3) or Concepts (Sec. 3.2.2)
3: SWM$[|\mathcal{P}|, k] \leftarrow$ TF-IDF $\times$ SM$^{\mathrm{T}}$

We apply this weighting approach in all three graph representations, i.e. Creator-oriented, Last Reply-oriented, and All Previous-oriented. Afterwards, we use the ADA algorithm [11] to find the influencers on the each OSN graph representation. We expect that using the topic/concept based filtering we can discover different sets of experts, which are complementary to the ones discovered by ADA when applied to a network created without this filtering.

3.5. Weighted graph construction

Algorithm 1 gives the baseline procedure to compute the Semantic Weighted Matrix that assigns an score to every post according to all concepts or topics considered for the network construction. First, the algorithm builds up the TF−IDF matrix according to Eq. (1). Secondly, it computes the semantic matrix SM according to topic or concept based text mining described in Sections 3.3 and 3.2.2, respectively. Finally, the semantic weighted matrixSWM is computed as the matrix multiplication between TF−IDF and SM. Algorithm 2 describe the construction of the graph representation according to the 'Creator' approach defined in the previous section, starting from the construction of the SWM. Also, according to the graph building strategy, the arc weight $a_{i,j}$ is increased when the message distance is greater or equal than the threshold $d_m(P_i, P_j) \geq \theta$.

Algorithm 2 Creator Network.

Require: $\mathcal{P}$ (Posts)
Ensure: Network $\mathcal{G}_c = (\mathcal{N}, \mathcal{A}_c)$
1: Build SWM according to Algorithm 1
2: Initialize $\mathcal{N} = \{\}, \mathcal{A}_c = \{\}$
3: **for** each $i \in \mathcal{P}$ **do**
4: $\quad \mathcal{N} \leftarrow \mathcal{N} \cup i$
5: **end for**
6: **for** each $i \in \mathcal{P}.creator$ **do**
7: $\quad$ **for** each $j \in \{i.replies\}, i \neq j$ **do**
8: $\quad\quad$ **if** $d_m(P_i, P_j) \geq \theta$ **then**
9: $\quad\quad\quad a_{i,j} \leftarrow a_{i,j} + 1$
10: $\quad\quad\quad \mathcal{A}_c \leftarrow \mathcal{A}_c \cup a_{i,j}$
11: $\quad\quad$ **end if**
12: $\quad$ **end for**
13: **end for**

4. Discovery of key-members on a real OSN

Computational experiments have been carried out over the data from a specific OSN: Plexilandia http://www.plexilandia.cl. We use information obtained from interviews with the OSN administrators to validate and complement our semantic based analysis of social network activity.

4.1. The case study social network

Plexilandia is an OSN created by a group of people who have a common interest in building artefacts for musical effects, amplifiers, and audio equipment in a "Do it yourself" style. It was born as a social network where to share common experiences in the construction of plexies[1]. Nowadays, Plexilandia has more than 2500 members after more than 9 years of existence. All these years they have been shearing and discussing their knowledge about building their own plexies, effects. Besides, there are other related topics such as luthier, professional audio, buy/sell parts that are covered by the activity of Plexilandia users. Users do not engage in conventional social (i.e. Facebook) relations. In the beginning the administration task was performed by only one member. Today, this task is performed by several administrators (in 2013 the OSN has 5 administrators) due to the increase in the amount of information generated weekly. Administration tasks are typically:

- Classification: it is frequent that social network members publish their contributions in the wrong forum section. For example, buy and sell advertisement should be placed in the "general forum" but newcomers place them in other sections. Therefore, administrators move them to the right place.
- Moderation: administrators must ensure that members use the forums to discuss topics which are related with the social network and language is appropriate (i.e. not offensive). This activity is carried out less frequently than classification, and requires less attention, since other active members help to detect these situations.
- Participation: although the social network knowledge is distributed all over its members, some members have greater degree of knowledge or expertise about some topics. Due to diverse reasons (being social network founders, experts in an area, etc.) administrators are significant knowledge generators. Therefore, they are active participants of most discussions. They start and stimulate new discussions and create new forums.

During its nine years of life, this OSN has undergone a sustained growth in members and their contributions. The analysis of the OSN behaviour by of administrators and experts about the

[1] "Plexi" is the nickname given to Marshall amp heads model 1959 that have the clear perspex (a.k.a plexiglass) fascia to the control panel with a gold backing sheet showing through as opposed to the metal plates of the later models.

Table 1
Plexilandia Activity measures.

Forum	2006	2007	2008	2009	2010	2011	2012	2013	2014	TOTAL
Amplifiers	392	2165	2884	3940	3444	3361	2398	1252	985	20822
Effects	184	1432	3362	3718	4268	5995	4738	2317	1331	27345
Luthier	34	388	849	1373	1340	2140	926	699	633	8382
General	76	403	855	1200	2880	5472	3737	1655	1295	17573
Pro Audio	–	–	–	–	–	342	624	396	219	1579
Synthesizers	–	–	–	–	–	–	–	104	92	196
TOTAL	686	4388	7950	10231	11932	17310	12423	6423	4555	75898

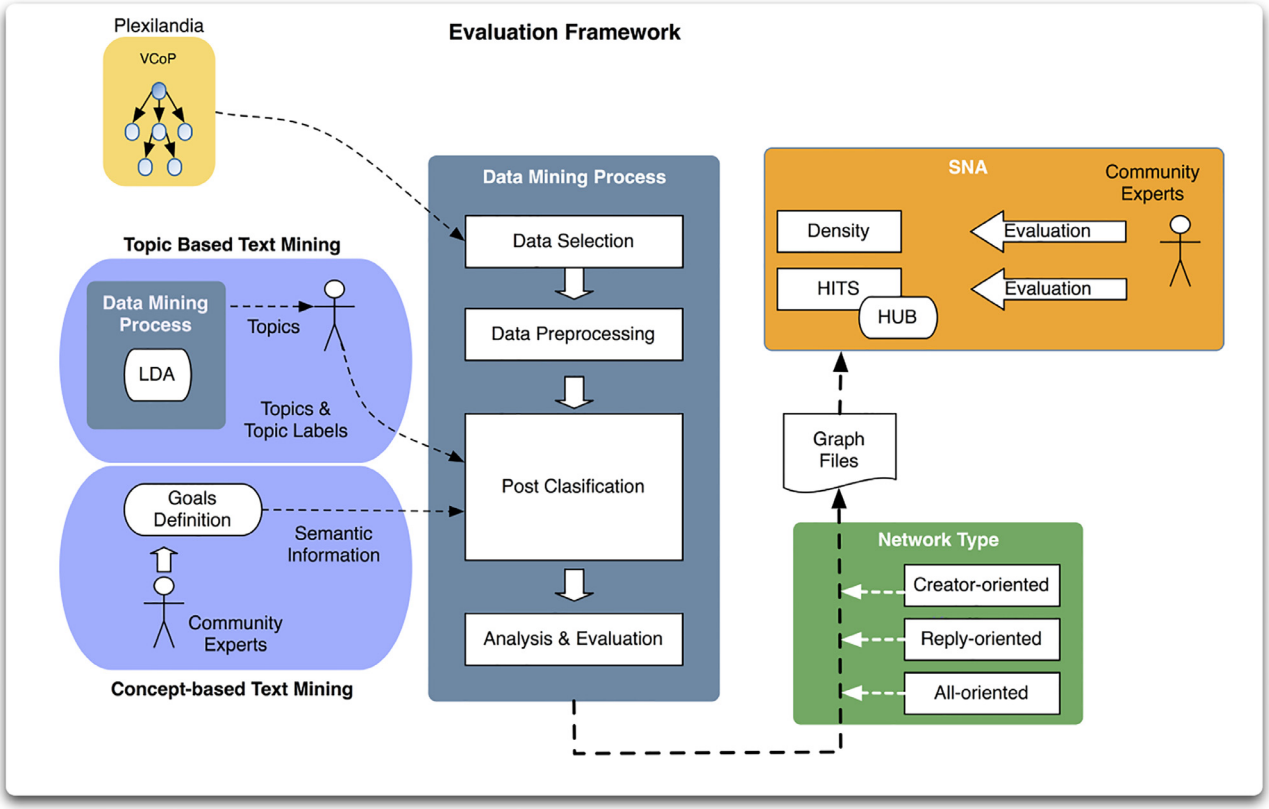

Fig. 3. Evaluation framework for fuzzy concept based and latent topic analysis methods.

social network is qualitative, grounded on their experience and time participating in the OSN. Besides, administrators have some global quantitative measures such as, for example, the total number of posts, the connected members, etc. (see Table 1 for a more exhaustive list). However, they don't have members browsing behaviour information, members publications quality, and how they contribute to the OSN purpose.

4.2. Evaluation framework

We expect the semantic concept and topic analysis to be able to detect more social network experts than the conventional approach of building the graph representation. In order to validate this semantic enriched search for the network key-members, we use the evaluation framework shown in Fig. 3. Firstly, we compute the density of the semantically filtered Creator graph representations $\mathcal{G}_c$ plus the contrast conventional graph without filtering. Secondly, we extract OSN key-members by the ADA algorithm, using ADA hub ranking criterion, to discover key-members according to the definition of key-members given above.

We have Plexilandia's forum data[2] since September 2006 until June 2014. However, the graph density study is carried out first only on whole years data from 2006 up to 2013. Later, we perform the same study for every month of year 2013. In the discovery of key-members, administrators have a hard time remembering to identify the experts of the social network after some years have passed, or even harder if you ask about the social network experts on a specific month. Fortunately, this kind of members is a stable cluster of users. Hence we will be using year 2013 in the evaluation of this work. Plexilandia network administrators provided a list of 64 key-members for year 2013 and 2014. We selected two from five administrators (the owners) and carried out questionnaires every two months in to identify the key members of the community on those two months. Often, once a key member is identified he or she maintain its expert status, the interesting case is when a new expert appear or when a member that belong to the core is passed to the key members group based on the administrators criteria. In the end, we found a total of 64 key members for 2013 and 2014, which are the ground truth for our evaluation.

[2] An important remark is that all data and results of our works are anonimized to avoid any privacy concern a social network member might have.

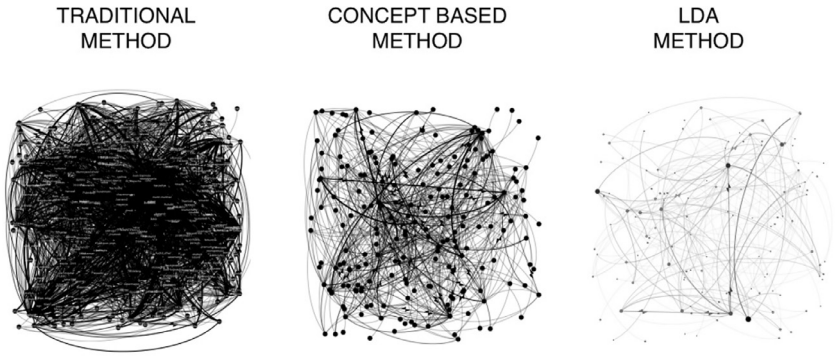

Fig. 4. Social Network graph density reduction with all methods for year 2013.

An interesting problem is that the OSN administrators were not able to rank all influencers as a key member, but they provide us with a refined list were they established sub categories of member importance as follows:

- Experts Type A: They are the most important key-members, 34 members have been identified for year 2013 based on administrators' criteria.
- Experts Type B: They are next in importance key-members. Administrators have identified 21 for the same period.
- Experts Type C: They are historic key-members, because they have been involved in the social network since its origins though they are not very active participating. Administrators have identified in this class 11 members.
- Experts Type X: these class is all members of the social network which are not key-members. They don't belong to the social network core and usually they ask questions rather than publish answers or tutorials.

4.3. Concept & topic based arc filtering

In this section we discuss the graph density reduction achieved from semantic arc filtering, either concept or topic based. We have depicted in Fig. 4 the Creator graph representations of Plexilandia for the whole year 2013, without filtering (traditional), and after semantic filtering by fuzzy concept and topic [LDA] arc filters. The threshold applied for both semantic filtering approaches is set to $\theta = 0.8$. Visual inspection shows that the semantic filtering achieves a remarkable thinning of the graph topology. In Figs. 5 and 6 we plot the evolution of the graph density per month in the years interval 2006–2013 and the year 2013, respectively. Each line plot corresponds to an arc filtering approach. From the perusal of these figures, it is easy to assess that concept based arc filtering always achieves a big density reduction relative to the non filtered graph. Moreover, the topic based arc filtering offers even greater density reduction. The effect of density reduction is that the computation of the key-members is easier and quicker. In fact, the ADA computational time is reduced proportionally to the number of arcs. Though we have not make precise quantitative measurements, we found that the cost of semantic filtering does not affect much this computational time reduction. In the next section we show that this simplification does not worsen the quality of key-member detection.

4.4. Influencer detection in graphs built by Creator strategy

The semantic filtering produces a strong reduction of graph density, with a strong reduction of the computational effort required for the extraction of key-members. In this section we examine how this computational simplification affects the quality of key-member extraction for the specific case of the graphs built applying the Creator strategy. We apply the ADA algorithm to the unfiltered graph (ADA), and to the graphs after the concept (ADA-FCA) and topic based (ADA-LDA) arc filtering, and we compare the discovered key-members with those provided by the OSN administrators, the ground truth for validation. We have not considered sequentially applying LDA after FCA or viceversa Table 2 gives the number of ground truth key-members discovered by each method, for 5 repetitions of the algorithm, using a fixed number of cut-off hubs 50 in the ADA algorithm. Because each approach discovers a different subset of the ground truth key-members, it is possible to improve the key-member accuracy by merging results from them. In Table 2 the entries ADA ∪ ADA-FCA and ∪ ADA-LDA denote the union of key-members discovered by the ADA algorithm over unfiltered and filtered graphs, which provide the best results for each set of key-members considered. We have illustrated this fact by Fig. 7, were we consider the ADA ranking over the unfiltered graph and the topics (LDA) filtered graph, setting the cut-off number of hubs to 10, 20, 30 and 40, looking for the discovery of A, A + B and A + B + C ground truth key-members, for the last five months of the year 2013. The middle part of each column (purple) is the intersection discovered by both methods. It can be appreciated that there is a core of key-members that is discovered by both approaches, but there are non-negligible subsets whose discovery is method dependent. The key-member ground truth was provided by the CSN administrators from a general consideration of the

Table 2
Number of key-members detected for different methods, setup of θ and ρ semantic filtering parameters, and cut off number of hubs in ADA algorithm set to 50. A, A+B, and A+B+C refers to set of ground truth key-member set considered. ADA, ADA-FCA, and ADA-LDA denote that the ADA is applied on the unfiltered graph, FCA and LDA arc filtered graphs, respectively. Columns correspond to the last five months in the year 2013. Bold highlights best results for each ground truth set.

Type of Expert	Method (θ/ρ)	1	2	3	4	5
A	ADA	18	16	20	18	13
	ADA-FCA (0.8/0.2)	16	12	20	14	13
	ADA ∪ ADA-FCA (0.8/0.2)	**22**	**19**	**23**	**18**	**16**
	ADA-LDA (0.4/0.01)	18	13	20	16	15
	ADA ∪ ADA-LDA (0.4/0.01)	**21**	**19**	**25**	**19**	**16**
A+B	ADA	26	25	29	27	23
	ADA-FACA (0.8/0.2)	23	20	28	24	21
	ADA ∪ ADA-FCA (0.8/0.2)	**32**	**29**	**34**	**30**	**26**
	ADA-LDA (0.4/0.01)	27	22	29	27	22
	ADA ∪ ADA-LDA (0.4/0.01)	**31**	**30**	**36**	**31**	**26**
A+B+C	ADA	31	29	33	29	26
	ADA-FCA (0.8/0.2)	27	25	34	26	23
	ADA ∪ ADA-FCA (0.8/0.2)	**38**	**34**	**40**	**33**	**29**
	ADA-LDA (0.4/0.01)	29	26	33	29	26
	ADA ∪ ADA-LDA (0.4/0.01)	**36**	**40**	**41**	**33**	**30**

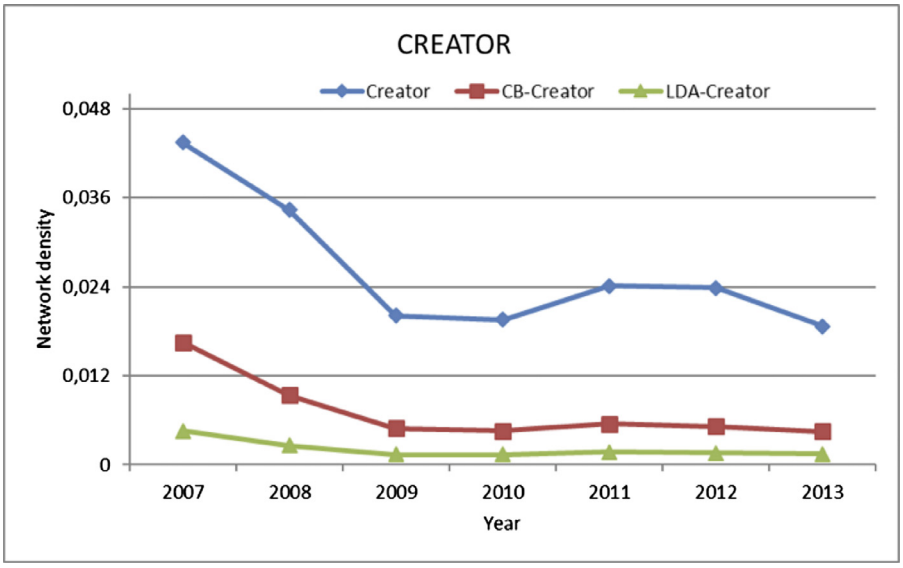

Fig. 5. Network density from years 2006 to 2014, we plot the densities of fuzzy concept (CB, red) arc filtering, latent topic (LDA, blue) arc filtering, and the non filtered graph (green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

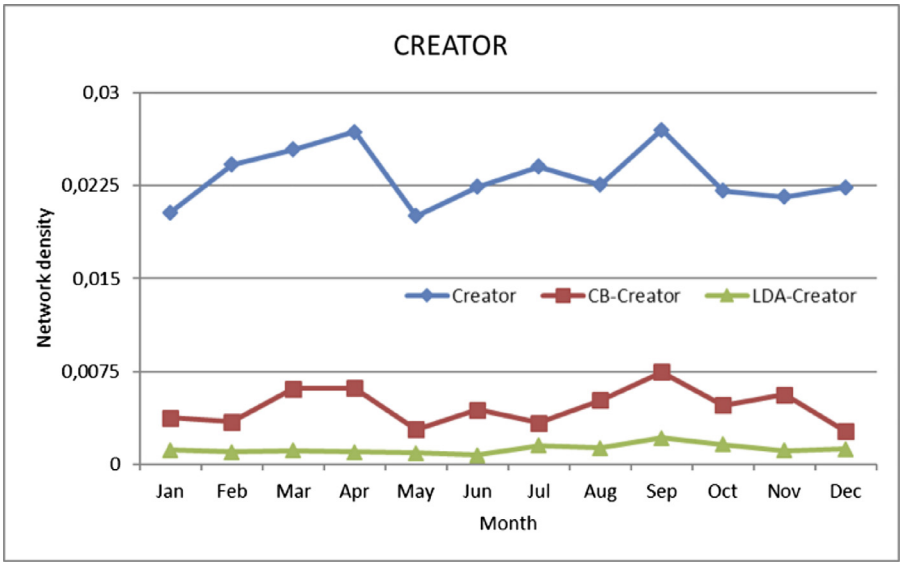

Fig. 6. Network density for year 2013, we plot the densities of fuzzy concept (CB, red) arc filtering, latent topic (LDA, blue) arc filtering, and the non filtered graph (green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

activity of the CSN during its life. The short time analysis carried out by the ADA on specific periods may reveal who was active in this period, so that we may consider the intersection as the core of the CSN key-members upon all methods agree, while other detections may correspond to lower activity key-members which may or may not be detected. Hence, fluctuations in key-member activity are detected by the detailed graph analysis which are not easy to validate through administrator experience.

We compute Precision, Recall and F-score for the key-members detected by ADA algorithm compared with the administrators ground truth. Tables 3–5 give these performance indicators for the three ground truth sets, respectively, specifically for the unfiltered and the LDA filtered graphs, with $\theta = 0.4$ and $\rho = 0.01$. Precision is greater on the unfiltered graph, but recall and F-score are higher for the combination of results, which is an improved detection of the key-members. It is also much more difficult to detect only the core key-members (type A) which suggests that their activity is less constant in time than assumed by the system administrators.

Table 3
Key-members detection precision, recall and F-score performance for Type *A* ground truth key-members of ADA applied to the unfiltered (ADA) and LDA arc filtered (ADA-LDA) graph. Columns correspond to the last five months in the year 2013.

Period	1	2	3	4	5
Precision ADA	0.45	0.40	0.50	0.45	0.33
Precision ADA ∪ ADA-LDA	0.41	0.35	0.48	0.38	0.31
Recall ADA	0.43	0.38	0.48	0.43	0.31
Recall ADA ∪ ADA-LDA	0.50	0.45	0.60	0.45	0.38
F-score ADA	0.44	0.39	0.49	0.44	0.32
F-score ADA ∪ ADA-LDA	0.45	0.39	0.53	0.41	0.34

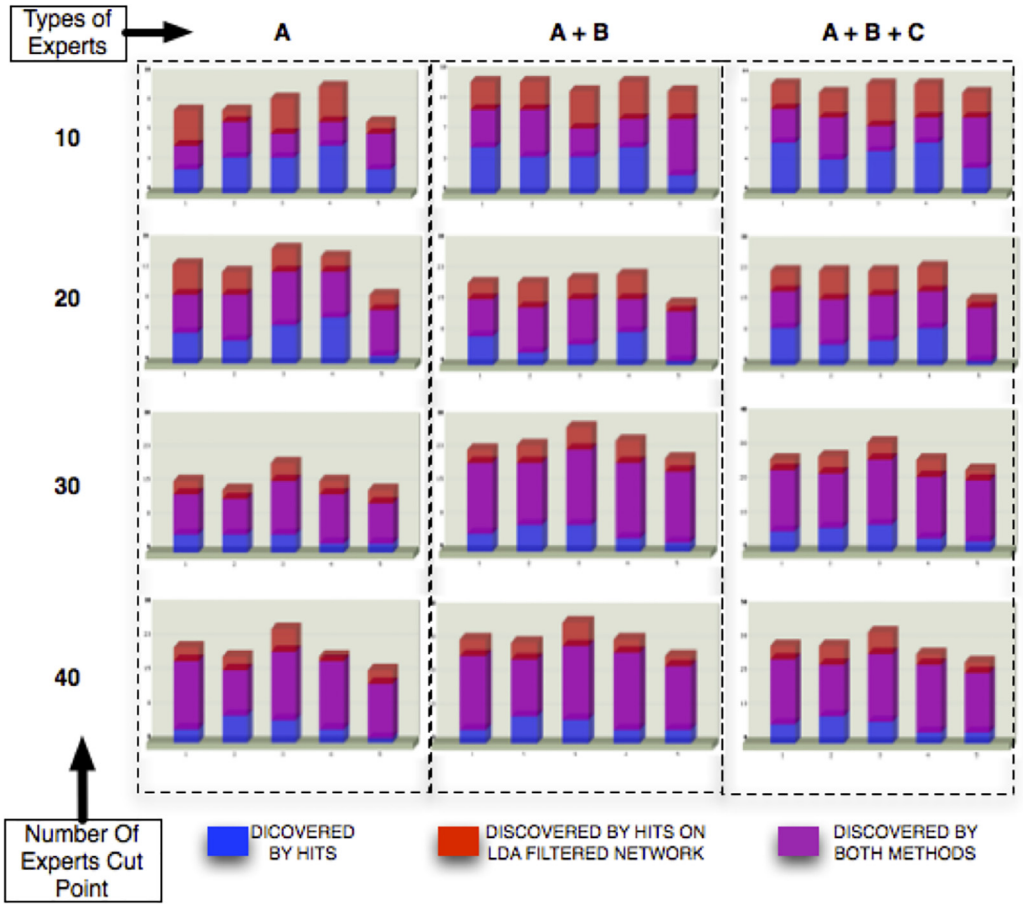

Fig. 7. Intersection of ground truth key-member discovered by ADA on the unfiltered and LDA filtered graphs, cut-off number of hubs set to 10, 20, 30, 40 members, for A, A+B, and A+B+C ground truth sets. Columns in the plots correspond to the months from August to December in the year 2013.

Table 4
Key-members detection precision, recall and F-score performance for Type $A+B$ ground truth key-members of ADA applied to the unfiltered (ADA) and LDA arc filtered (ADA-LDA) graph. Columns correspond to the last five months in the year 2013.

Period	1	2	3	4	5
Precision ADA	0.65	0.63	0.73	0.68	0.58
Precision ADA ∪ ADA-LDA	0.61	0.55	0.69	0.62	0.50
Recall ADA	0.47	0.45	0.53	0.49	0.42
Recall ADA ∪ ADA-LDA	0.56	0.55	0.65	0.56	0.47
F-score ADA	0.55	0.53	0.61	0.57	0.48
F ADA ∪ ADA-LDA	0.58	0.55	0.67	0.59	0.49

Table 5
Key-members detection precision, recall and F-score performance for Type $A+B+C$ ground truth key-members of ADA applied to the unfiltered (ADA) and LDA arc filtered (ADA-LDA) graph. Columns correspond to the last five months in the year 2013.

Period	1	2	3	4	5
Precision ADA	0.78	0.73	0.83	0.73	0.65
Precision ADA ∪ ADA-LDA	0.71	0.65	0.79	0.66	0.58
Recall ADA	0.48	0.45	0.51	0.45	0.40
Recall ADA ∪ ADA-LDA	0.55	0.55	0.63	0.51	0.46
F-score ADA	0.59	0.55	0.63	0.55	0.50
F-score ADA ∪ ADA-LDA	0.62	0.60	0.70	0.57	0.51

5. Conclusion

As web services are evolving into social networks social analysis tools are becoming part of the toolbox of system administrators. When considering online social networks supporting a community of practice, with a narrow common interest, one of the most important tasks in order to assess the healthy status of the network is the identification of the members that contribute with ideas and projects, i.e. the key-members, influencer or experts. In this paper, we aim to complement the influencer identification algorithms with a semantic filtering of the graph representation arcs, expecting that such a simplification would reduce the computational cost of influencer detection without degradation of the quality of the detection. We proposed two kinds of simplification, one based on fuzzy concept modelling and the other on topics discovered by LDA approach. The later produced a greater simplification. Both semantic graph simplification approaches did not degrade the influencer detection, when compared with static identification ground truth provided by the system administrators in a specific community of practice online social network. The semantic simplification lead to discover complementary members, enhancing the baseline discovery over the unfiltered graph. Regarding scalability, the proposed approach improves the scalability of the underlying key-member detection algorithm, whose complexity is directly proportional to the number of nodes and arcs, because it requires to update and revisit all the node weights. Semantic analysis on the other hand has strong computational requirements. Some specific implementations, such as variational inference for LDA, allow to work with large datasets, but to scale the approach to the size magnitudes

of current social networks, efficient distributed implementations would be required in order to speed-up the algorithms

5.1. Future work

A higher order semantic analysis, via the automatic ontology creation from the posted documents, would be desirable. These ontologies would allow to define more precise semantic distances, as distances between concepts in the structure of the ontology, with would allow finer analysis of the relations between users, and the strict relevance of the interactions to the purpose of the social networks. The ADA authority ranking may be complemented with other SNA metrics such as influence spread, centrality, and others. In the long term, our works would derive in administration tools that may allow a real time visualisation of the social network status, in order to take appropriate policies to ensure its healthy activity. The application of the approach to conventional online social networks, i.e. those that do not have a well defined content orientation, requires an extension of the topic modelling from text based analysis to multimedia contents in order to apply semantic filtering. Such extension is currently underway in the research community with the application of deep learning approaches able to extract semantic meaning from image and sound data.

Acknowledgement

Support from the Chilean "Instituto Sistemas Complejos de Ingeniera" (ICM: P-05-004-F, CONICYT: FBO16; www.sistemasdeingenieria.cl) and the Business Intelligence Research Center (www.ceine.cl).

This work wouldn't be possible without Plexilandia, therefore, we would like to thank Plexilandia Community. Specially, Mr. Jose Ignacio Santa-Cruz, who has provided always the right information; besides, his good observations on the day-to-day community's behaviour.

References

[1] H. Alvarez, S.A. R'ios, F. Aguilera, E. Merlo, L. Guerrero, R. Setchi, et al., Enhancing social network analysis with a concept-based text mining approach to discover key members on a virtual community of practice, in: (Eds.), Knowledge-based and Intelligent Information and Engineering Systems, vol. 6277, Springer-Verlag, Heidelberg Ge., 2010, pp. 591–600.
[2] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[3] A. Bourhis, L. Dubé, R. Jacob, The success of virtual communities of practice: The leadership factor, Electronoc J. Knowl. Manag. 3 (2005) 23–34.
[4] C. Chen, S. Hung, To give or to receive? factors influencing members' knowledge sharing and community promotion in professional virtual communities, Inform. Manag. 47 (2010) 226–236.
[5] Y. Fang, C. Chiu, In justice we trust: Exploring knowledge-sharing continuance intentions in virtual communities of practice, Comput. Human Behav. 26 (2010) 235–246.
[6] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (2010) 75–174.
[7] L.C. Freeman, Centrality in social networks conceptual clarification, Social Netw. 1 (1978) 215–239.
[8] J. Gairín-Sallán, D. Rodríguez-Gómez, C. Armengol-Asparó, Who exactly is the moderator? a consideration of online knowledge management network moderation in educational organisations, Comput. Educ. 55 (2010) 304–312.
[9] C. Johnson, A survey of current research on online communities of practice, Int. Higher Educ. 4 (1) (2001) 45–60.
[10] A.J. Kim, Community Building on the Web: Secret Strategies for Successful Online Communities, Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 2000.
[11] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, J. ACM 46 (1999) 604–632.
[12] M. Kosonen, Knowledge sharing in virtual communities – a review of the empirical research, Int. J. Web Based Commun. 5 (2009) 144–163.
[13] H. Kwak, Y. Choi, Y.H. Eom, H. Jeong, S. Moon, Mining communities in networks: a solution for consistency and its evaluation, in: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference, pp. 301–314.
[14] G. LHúillier, H. Alvarez, S.A. Ríos, F. Aguilera, Topic-based social network analysis for virtual communities of interests in the dark web, in: Proceedings of the 16th International Conference on Knowledge Discovery on Databases (KDD), 2010.
[15] X. Liu, W.B. Croft, M. Koll, Finding experts in community-based question-answering services, in: Proceedings of the CIKM'05, 2005, pp. 315–316.
[16] A. McCallum, A. Corrada-Emmanuel, X. Wang, Topic and role discovery in social networks, in: Proceedings of the IJCAI, 2005, pp. 786–791.
[17] A. McCallum, X. Wang, A. Corrada-Emmanuel, Topic and role discovery in social networks with experiments on enron and academic email, J. Artif. Intell. Res. 30 (2007) 249-272.
[18] E. Merlo, S. Ríos, H. Álvarez, G. LHúillier, J. Velásquez, R. Setchi, Finding inner copy communities using social network analysis, in: Knowledge-based and Intelligent Information and Engineering Systems, LNAI, vol. 6277, Springer-Verlag, Heidelberg, Ge, 2010, pp. 581–590.
[19] H. Nakanishi, I. Turksen, M. SUGENO, A review and comparison of six reasoning methods, Fuzzy Sets Syst. 57 (3) (1993) 257-294.
[20] R.D. Nolker, L. Zhou, Social computing and weighting to identify member roles in online communities, in: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, 2005, pp. 87–93.
[21] N. Pathak, C. Delong, A. Banerjee, K. Erickson, Social topic models for community extraction, in: Proceedings of the 2nd SNA-KDD Workshop ?08(SNA-KDD?08), 2005.
[22] U. Pfeil, P. Zaphiris, Investigating social network patterns within an empathic online community for older people, Comput. Human Behav. 25 (2009) 1139–1155.
[23] X.H. Phang, C. Nguyen, Gibbslda++, https://github.com/mrquincle/gibbs-lda, http://gibbslda.sourceforge.net/, last view 08/16/2016.
[24] C.E. Porter, A typology of virtual communities: A multi-disciplinary foundation for future research, J. Comput.-Med. Commun. 10 (2004) 00.
[25] G. Probst, S. Borzillo, Why communities of practice succeed and why they fail, Eur. Manag. J. 26 (2008) 335–347.
[26] S.A. Ríos, F. Aguilera, 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, 5712, 2009, pp. 480–489.
[27] S.A. Ríos, J.D. Velásquez, H. Yasuda, T. Aoki, A hybrid system for concept-based web usage mining, Int. J. Hybrid Intell. Syst. 3 (2006) 219–235.
[28] G. Salton, A. Wong, C.S. Yang, A vector space model for automatic indexing, Commun. ACM 18 (1975) 613–620.
[29] S. Toral, M. Martínez-Torres, F. Barrero, Analysis of virtual communities supporting OSS projects using social network analysis, Inform. Softw. Technol. 52 (2010) 296–303.
[30] K.D. Valck, G.H. van Bruggen, B. Wierenga, Virtual communities: A marketing perspective, Dec. Support Syst. 47 (2009) 185–203.
[31] S. Wasserman, K. Faust, Social Network Analysis: Methods and Applications, Cambirdge University Press, Cambridge, UK, 1995.
[32] B. Wellman, Computer networks as social networks, Science 293 (2001) 2031–2035.
[33] B. Wellman, M. Gulia, Virtual communities as communities, in: M.A Smith, P. Kollock (Eds.), Communities in Cyberspace, Routledge, Oxford, UK, 1999.
[34] B. Wellman, J. Salaff, D. Dimitrova, L. Garton, Computer networks as social networks: collaborative work, telework, and virtual community, Annual Rev. Sociol. Vol. 22 (1996) 213–238.
[35] D. Xing, M. Girolami, Employing latent Dirichlet allocation for fraud detection in telecommunications, Pattern Recogn. Lett. 28 (2007) 1727–1734.
[36] J. Xu, H. Chen, Crimenet explorer: a framework for criminal network knowledge discovery, ACM Trans. Inform. Syst. 23 (2) (2005) 201–226.
[37] L. Yang, F. Liu, J. Kizza, R. Ege, Discovering topics from dark websites, in: Proceedings of the IEEE Symposium on Computational Intelligence in Cyber Security, 2009, pp. 175–179.
[38] K. Yelupula, S. Ramaswamy, Social network analysis for email classification, in: Proceedings of the 46th Annual Southeast Regional Conference, 2008, pp. 469–474.

Sebastián A. Ríos is Assistant Professor at the Industrial Engineering Department of the University of Chile since (2008). He received the B.E on Industrial Engineering on 2001, the B.E on Computer Science, P.E. on Industrial Engineering on 2003 from the University of Chile, Chile; and the Ph.D. on Knowledge Engineering from the University of Tokyo, Japan on 2007. He is the Founder and Director of the Business Intelligence Research Center (CEINE) at the University of Chile since 2012, a collaborative applied research effort between private companies and the University. His research interests include data mining algorithms in big dataset and its applications to different industry domains (medicine, marketing, management, etc.); he also is interested in generative topic models for text mining in social networks and knowledge representation using semantic web technologies.

J. David Nuñez-Gonzalez is Ph.D. student and researcher at University of the Basque Country (UPV/EHU). Is member of Computational Intelligent Group whose responsible is Prof. Manuel Graña, being the advisor of Ph.D. too. In this short research period, author has already contribute to five international conferences and is working in Social and Smart European Project coordinated by Prof. Bruno Apolloni (University of Milan). His current research interests are data mining, machine learning, social networks, trust computing, recommendation systems, ant colony optimization and harmonic search.

Manuel Graña Romay received the M.Sc. and Ph.D. degrees from Universidad del Pais Vasco (UPV/EHU), Donostia, Spain, in 1982 and 1989, respectively, both in computer science. His current position is a Full Professor (Catedrático de Universidad) with the Computer Science and Artificial Intelligence Department of the Universidad del Pais Vasco (UPV/EHU). He is the head of the Computational Intelligence Group (Grupo de Inteligencia Computational). His current research interests are in applications of computational intelligence to linked multicomponent robotic systems, medical image in the neurosciences, multimodal human computer interaction, remote sensing image processing, content based image retrieval, lattice computing, semantic modelling, data processing, classification, and data mining.