A New Content-Based Model for Social Network Analysis

Paola Velardi(*), Roberto Navigli (*), Alessandro Cucchiarelli (**), Fulvio D'Antonio (**)
(*)University of Roma "La Sapienza", (**)Polytechnic University of Marche
{velardi,navigli}@di.uniroma1.it, {cucchiarelli,dantonio}@diiga.univpm.it

Abstract

This paper presents a novel model for social network analysis in which, rather than analyzing the quantity of relationships (co-authorships, business relations, friendship, etc.), we analyze their communicative content. Text mining and clustering techniques are used to capture the content of communication and to identify the most popular themes. The social analyst is then able to perform a study of the network evolution in terms of the relevant themes of collaboration, the detection of new concepts gaining popularity, and the existence of popular themes that could benefit from better cooperation.

The methodology is experimented in the domain of a Network of Excellence on enterprise interoperability, INTEROP.

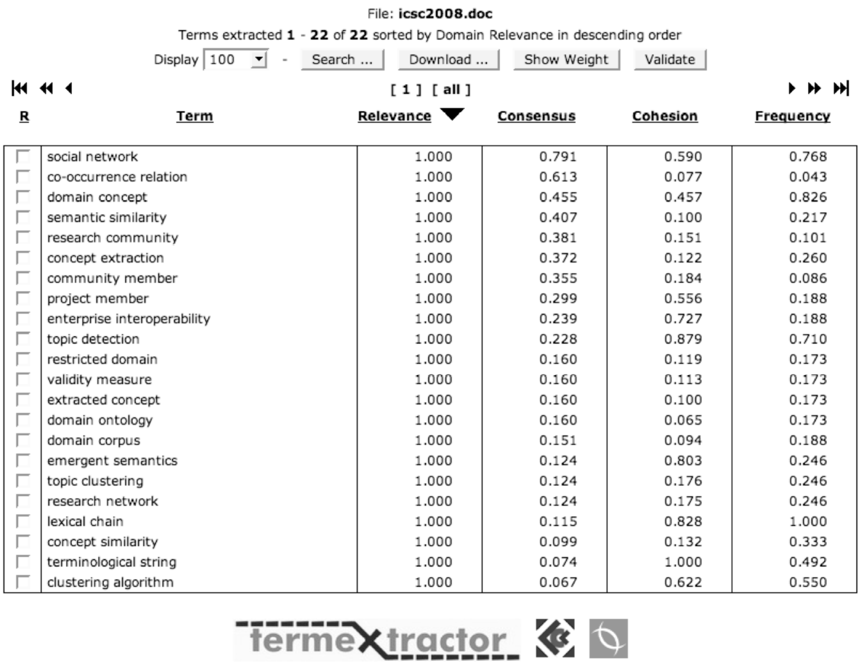
1. Introduction

Relationships among actors in traditional social network analysis are modeled as a function of the *quantity* of relations (co-authorships, business relations, friendship, etc.). In contrast, in a business, social or research community, network analysts are interested in the *communicative content* exchanged by the community members, not just in the number of relationships. To meet this need, this paper presents a novel social network model, named *content-based social networks analysis* (CB-SNA), in which actors are not simply represented through the intensity of their mutual relationships, but also through the analysis and evolution of their shared interests. The idea of CB-SNA is not entirely new: a method to discover groups of people sharing the same discussion topics in Google news [4], as well as the use of a Bayesian Network that captures topics and the directed social network of senders and recipients in a message-exchange environment [12] have been proposed. However, both approaches extract content with the aid of a naïf bag-of-words model. In contrast, we perform CB-SNA with the aid of deep linguistic analysis in three steps:

1. Concept extraction: the first step implies the extraction of the concepts that play a relevant role in describing actors' relationships. This is performed by collecting the written communications (e-mail, blogs, co-authored papers, documents of any type) exchanged among the community members, and extracting the set of relevant, domain-specific, concepts.
2. Topic detection: semantic similarity between concept pairs is computed by analyzing ontological and co-occurrence relations in the domain. A clustering algorithm is then used to group concepts into *topics*. A topic is *a set of semantically close concepts*, but the relevance of a topic is tied to the specific set of documents that characterize inter-actors communications in a given time interval (e.g. the scientific publications produced by a research community in a given time span). As a result of this step, a social relation between two actors (persons or organizations) can be modelled in terms of the involved topics.
3. Social Network Analysis: social network measures are used to model the evolution of collaboration content across time. A collaboration strength is usually computed ([9], [14]) as the number of common activities between pairs of actors (e.g. the number of common papers in a research community). In contrast, we weight a link between actors as a function of *topic overlapping*. This network is what we call content-based social network model (CB-SN). Traditional social network measures are adapted to the study of a CB-SN, in order to analyze the dynamics of the agents' aggregation around the topics.

The paper is organized as follows: Sections 2 and 3 describe in detail the methodology and algorithms used for concept extraction, topic detection and clustering analysis. Section 4 presents the CB-SN network model and measures. Finally, Section 5 is dedicated to concluding remarks and presentation of future work.

The CB-SN model is applied to the study of a research community in the field of enterprise interoperability, the INTEROP NoE. INTEROP (http://www.interop-vlab.eu) is a recently concluded Network of Excellence, now continuing its mission under the name of "Virtual Laboratory on Enterprise Interoperability".



File: **icsc2008.doc**

Terms extracted **1** - **22** of **22** sorted by Domain Relevance in descending order

Display 100 ▼ - Search ... Download ... Show Weight Validate
⏮ ⏪ ◀ [1] [all] ▶ ⏩ ⏭
termeXtractor

R	Term	Relevance ▼	Consensus	Cohesion	Frequency
□	social network	1.000	0.791	0.590	0.768
□	co-occurrence relation	1.000	0.613	0.077	0.043
□	domain concept	1.000	0.455	0.457	0.826
□	semantic similarity	1.000	0.407	0.100	0.217
□	research community	1.000	0.381	0.151	0.101
□	concept extraction	1.000	0.372	0.122	0.260
□	community member	1.000	0.355	0.184	0.086
□	project member	1.000	0.299	0.556	0.188
□	enterprise interoperability	1.000	0.239	0.727	0.188
□	topic detection	1.000	0.228	0.879	0.710
□	restricted domain	1.000	0.160	0.119	0.173
□	validity measure	1.000	0.160	0.113	0.173
□	extracted concept	1.000	0.160	0.100	0.173
□	domain ontology	1.000	0.160	0.065	0.173
□	domain corpus	1.000	0.151	0.094	0.188
□	emergent semantics	1.000	0.124	0.803	0.246
□	topic clustering	1.000	0.124	0.176	0.246
□	research network	1.000	0.124	0.175	0.246
□	lexical chain	1.000	0.115	0.828	1.000
□	concept similarity	1.000	0.099	0.132	0.333
□	terminological string	1.000	0.074	1.000	0.492
□	clustering algorithm	1.000	0.067	0.622	0.550

Figure 1. Terminology Extraction result from a draft version of this paper.

2 Concept extraction and concept analysis

The objective of this phase is to identify the *emergent semantics* of a community, i.e. the concepts that better characterize the *content* of actor's communications. Concepts are extracted from available texts (hereafter referred to as the *domain corpus*) exchanged among the members of the community.

Once concepts have been identified, a similarity metric is defined to weight relations between concept pairs. Conceptual relations are computed using the method of lexical chains [7], where chains are derived from three sources: co-occurrence data extracted from the domain corpus, concept glosses, and a domain ontology or thesaurus. We first describe the concept extraction methodology, then the algorithm to compute concept similarity.

2.1 Concept identification

It has often been pointed out ([10], [13], and [6]) that terminological strings (e.g. multi word sequences, or *key phrases*) are more informative features than single words for representing the content of a document. To extract terminological strings, we used the TermExtractor terminology extraction system [18], a freely accessible web application that we have developed. One of the relevant features of TermExtractor is that, unlike other terminology tools, it is able to extract the terms from an entire corpus, rather than from a single document (though single-document analysis is possible). TermExtractor selects the concepts that are *consistently* and *specifically* used across the corpus documents, according to information-theoretic measures, statistical filters and structural analysis. Several options are available to fit the needs of specific domains of analysis, among which, singleton terms extraction, named entity analysis and single user or group validation. As an example[1], Fig. 1 is a screenshot of the highest ranked terms extracted from a draft version of this paper.

2.2 Defining a semantic similarity measure

In natural language-related applications, similarity between words is measured either through statistical measures or by using a taxonomy or thesaurus (e.g. [16], [2], [6], [15], [20] and [23]). Similarity measures clearly benefit from the availability of ontologies or taxonomies: for example, knowing that *design process integration* is a hyponym of *business process* allows it to draw a similarity link between these two concepts, that would otherwise be undetectable with co-occurrence analysis, since hyponyms rarely co-occur. The use of taxonomies has so far been limited by two factors: in general domains, where large taxonomies are available (e.g. WordNet), semantic

[1] The interested reader may easily replicate the experiment by accessing http://lcl.uniroma1.it/termextractor.

similarity computation must cope with the complex problem of sense ambiguity. In restricted domains, where ambiguity is usually limited [5], high-coverage taxonomies are not available, with the exception of few domains, like medicine and art. Recently, we developed a methodology, named OntoLearn, to create a high-coverage domain ontology with limited manual effort, starting with a small, manually acquired core ontology. This technique was applied in the INTEROP project and led to the semi-automatic creation of an ontology on enterprise interoperability [21], which was evaluated in the large by the project members.

We based the similarity measure computation on a combination of co-occurrence data and ontology-based semantic relatedness, as detailed hereafter.

First, a graph $G=(V,E)$ is built, being V the set of nodes representing terminological strings (hereafter denoted also as domain concepts[2]) extracted as described in Section 2.1, and E the set of edges. An edge (t_j, t_i) is added to E if any of the following three conditions hold:

i) A relation holds between the concepts expressed by t_j and t_i in the ontology (e.g. *ontology representation* is a kind-of *knowledge representation*). Note that edges are directed;

ii) the term t_i occurs in a textual definition of t_j from the domain glossary (e.g. we add the edge (*ontology representation, ontology*) to E, as *ontology representation* is defined as "the description of an *ontology* in a well-defined language");

iii) the two terms co-occur in the document corpus according to the Dice coefficient:

$$Dice(t_j,t_i) = \frac{f(t_j,t_i)}{f(t_j)+f(t_i)}$$

where $f(t_j)$ and $f(t_i)$ are the occurrence frequency of t_j and t_i in the document corpus, and $f(t_j, t_i)$ is the co-occurrence frequency of the two terms (e.g. we add to E the edge (*ontology representation, ontology model*)).

Experiments in this paper are conducted using the INTEROP document repository (a set of research papers collected by the project members) and the INTEROP ontology, however the availability of an ontology is not fully critical, since sufficiently rich information can be obtained using the document corpus and a domain glossary. In Figure 2 we show an excerpt of the graph G obtained as described above.

Given the graph G, for each pair of concepts t_j and t_i, we compute the set of lexical chains in the graph, i.e. edge paths of length l ($l = 1, ..., L$, where L is the maximum path length) which connect the two concepts:

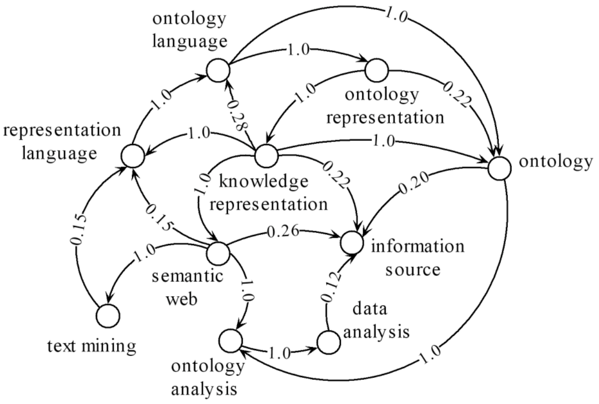


Figure 2. An excerpt of the graph built for computing the lexical chains.

$$LC_l(t_j,t_i) = \{t_j \equiv t_1 \rightarrow t_2 \rightarrow ... \rightarrow t_{l-1} \rightarrow t_l \equiv t_i\}$$

Finally, we compute the semantic similarity between t_j and t_i as a function of the corresponding lexical chains between the two concepts:

$$sim(t_j,t_i) = \sum_{l=1}^{L} \frac{|LC_l(t_j,t_i)|}{|LC_l(t_j)|} e^{-l}$$

where $LC_l(t_j)$ denotes the set of all the lexical chains connecting t_j to any other node (i.e. the union of the sets $LC_l(t_j, t_m)$ for all $t_m \in V \backslash \{t_j\}$). According to the above formula, the contribution of the lexical chains of length l is given by the inverse of the exponential of l weighted by the ratio of lexical chains of length l which connect t_j to t_i to that which connect t_j to any node in the graph.

Each domain concept t_j is then associated with an n-dimensional vector x_j, where n is the total number of extracted concepts, and the k-th component of x_j is $x_{ji}=sim(t_j,t_i)$. In the following, we denote with X the space of instance vectors, where $|X|=|V|=n$.

3. Topic detection

In the previous section we introduced a novel methodology to identify domain concepts and compute concept similarity vectors, which is based on automated terminology extraction and ontology-enriched lexical chains analysis.

The subsequent step, *topic detection*, is a *clustering* task: the objective is to organize concepts in groups, or clusters, so that concepts within a group are more similar to each other than are concepts belonging to different clusters.

The literature of clustering methods is vast (see [8] and [19] for a survey), but even in recent studies [3] *k-means* seems to compete with other methods for

[2] Words and concepts have a many to many correspondence because of ambiguity and synonyms. However, in restricted domains, we might assume a one to one relation between terminological strings and domain concepts.

simplicity and accuracy of results. We used an empowered version of the algorithm, the *k-means++* method for optimal selection of the initial seeds [1]. The authors show that, by augmenting k-means with a randomized seeding technique, they obtain an algorithm that is *$\Theta(logk)$*-competitive with optimal clustering. To evaluate alternative clustering results, we used an unsupervised validity measure called the *Silhouette Coefficient*[3] a popular method that combines the notions of inter-cluster *cohesion* and intra-cluster *separation*.

3.1 The clustering algorithm

The clustering algorithm is the following:

1. Heuristically determine[4] a set of integer values $K=\{k_1,k_2,...,k_p\}$ where k_i is the number of clusters to be acquired.
2. For each $k_i \in K$ do:
   - Select *k_i-best* initial cluster centers $V^{0,ki}=\{x_1,x_2,..., x_{ki}\}$, according to the *k-means++* method;
   - Run *k-Means*$(k_i, V^{0,ki})$[5]. Let $\mathcal{C}(k_i)$ be the final classification of vectors in X when $k=k_i$ and let $\{C_1,C_2,..., C_{ki}\}$ be the resulting clusters.
   - For any $x_j \in X$ and $C_h \in \mathcal{C}(k_i)$, define the distance between x_j and C_h as:

     $$d(x_j,C_h) = \sum_{x_k \in Ch} \frac{d(x_j,x_k)}{|C_h|}$$ where d is the euclidean distance

     $a(x_j,\mathcal{C}(k_i)) = d(x_j,C_m)$ if $x_j \in C_m$, and

     $$b(x_j,\mathcal{C}(k_i)) = \min_{\substack{C_h \in \mathcal{C}(k_i), \\ C_h \neq C_m}} \left(d(x_j,C_h)\right)$$
   - Compute the Silhouette coefficient as:

     $$S(x_j,\mathcal{C}(k_i)) = \frac{b-a}{\max(a,b)} \text{ and } S(\mathcal{C}(k_i)) = \frac{\sum_{x_j \in X} S(x_j,\mathcal{C}(k_i))}{|X|}$$

     To reduce the high contribution of singleton clusters to the Silhouette value[6], we *smoothed* the Silhouette formula as:

     $$S'(C_h) = \frac{1}{|C_h|} f(C_h,|X|) \sum_{x_j \in C_h} S(x_j,\mathcal{C}(k_i))$$

     and $S'(\mathcal{C}(k_i)) = \frac{\sum_{C_h \in \mathcal{C}(k_i)} \cdot S'(C_h)}{|k_i|}$,where *$f(C_h,|X|)$* is a function that decreases with the ratio $C_h / |X|$[7].
3. Let $\mathcal{C}(k_{BEST})$ be the best classification obtained according to the adopted validity measure S.

Hereafter we simply indicate with $\mathcal{C}=\{C_1,C_2, ..., C_k\}$ the best clustering result. Each C_h in $\mathcal{C}$ represents a *topic*, i.e. a subset of highly semantically correlated concepts, modelling the communication between actors in a social network.

3.2 Experiments on topic clustering

We used the methodology described in Sections 2 and 3 to extract the relevant research topics of the INTEROP community.

We collected 1452 full papers or abstracts authored by the INTEROP project members belonging to 46 organizations. Table 1 summarizes the relevant results and data of the corpus analysis phase.

Table 1. Summary data on corpus analysis

Number of analyzed papers	1452
Extracted terms	728
Domain Ontology used	http://interop-vlab.eu/backoffice/tav

The lexical chain methodology was then applied to the extracted concepts, using the semantic relations encoded in the INTEROP ontology, and the co-occurrence relations extracted from the domain corpus and from the INTEROP glossary.

An example of similarity vector (in which we show only the highest-rated arguments) is:

```
activity_diagram = (class_diagram (1),
process_analysis (0.630), software_engineering
(0.493), enterprise_software (0.488),
deployment_diagram (0.468), bpms_paradigm
(0.467), workflow_model (0.444), model-
driven_architecture (0.442),
workflow_management (0.418))
```

Finally, the concept vectors built from lexical chains were used to feed the k-mean++ algorithm. The cluster validity measure was computed for incremented values of *k*, *$50 \leq k \leq 300$*. Clustering results in the range *$140 \leq k \leq 170$* show the best S' values. Figure 3 shows the two best-rated cluster (using the *arctan(x)* smoothing of the cluster's Silhouette and *k*=150).

There are no straightforward ways for an objective evaluation of clustering results: external evaluation criteria (i.e. evaluation on standard datasets) are not possible, since no benchmarks are available on term

[3] www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf.

[4] E.g: http://nlp.stanford.edu/IR-book/html/htmledition/cluster-cardinality-in-k-means-1.html.

[5] We used the *weka* www.cs.waikato.ac.nz/ml/weka/ k-Means, modified with the ++ seeding methodology.

[6] Most cluster validity measures have an undesired behaviour at the extremes, e.g. when $k_i \to 1$ *or* $k_i \to |X|$.

[7] We experimented several functions with comparable results, e.g. based on *arctan(x)*, *x^2* etc.

```
cluster 19 = { common_ontology,
core_domain_ontology, core_ontology,
domain_ontology, enterprise_ontology,
federated_ontology, ontology_alignment,
ontology_analysis, ontology_application,
ontology_architecture, ontology_maintenance,
ontology_mediation, ontology_merging,
ontology_representation, ontology_validation,
ontology_versioning, reference_ontology}

cluster 8 = { application_interoperability,
enterprise-wide_network,
enterprise_interoperability,
information_interoperability,
interoperability_service, model_interoperability,
organisation_interoperability,
process_interoperability, service_interoperability,
system_interoperability}

cluster 89 = { automatic_composition,
b2b_connectivity,business_performance,business_
process_support,collaborative_task,component-
based_system,component_interaction,composition_
synthesis,distributed_architecture,distributed_
workflow_system,.. }
```

Figure 3. The 3 best clusters obtained with k=150.

clustering[8]. As far as internal evaluation criteria are concerned, it has been experimentally shown that none of the proposed validity indices reliably identifies the best clusters, unless these are clearly separated [11].

We then performed a qualitative analysis of the data, based on our experience and knowledge of the INTEROP community and research domains. Inspecting the data, the relevant phenomena in the range $140 \leq k \leq 170$ remain more or less the same:

- The "central" research themes of the INTEROP community constantly emerge: for example, it is always possible to find an "*ontology*" topic (like cluster 19 in Fig. 2), but, as k grows, an initially large cluster is split into more fine-grained sub-topics. For example, for k=150 there are 4 high-ranked ontology-related clusters. A similar behaviour is observed for *interoperability* (cluster 8), *business* and other relevant INTEROP themes.
- Roughly 20-25% of the concepts aggregate in a rather variable manner, eventually contributing to singleton clusters as k grows. This was expected, since, in natural language applications, a certain degree of data sparseness is indeed unavoidable, and even predictable[9]. However, to the extent that a significant number of relevant topics clearly emerge, this phenomenon does not affect the subsequent social network analysis.

4. Content-based analysis of social networks

The following section is dedicated to the description and analysis of a Content Based-Social Network. We refer to the specific case of a *research network*, but the approach is general whilst written material is available to model actor's relationships.

4.1 The content-based social network

Given the set $G = \{g_1, \ldots, g_{|C|}\}$ of the INTEROP research groups, the set D of the project members' publications (as described in section 3.2) and the collection V of *domain concepts* (as defined in section 2.2), a pattern-matching algorithm is used to tag each publication d_i in D with a subset of domain concepts $V_i \subseteq V$.

For any document d_i, we compute a vector v_i of k elements y_{ih} (with $k=|\mathcal{C}|$) such that:

$$y_{ih} = \frac{l_{h,i}}{|C_h|} \sum_{j:x_j \in C_h} tf \bullet idf(t_j, d_i)$$

where x_j is the similarity vector associated with concept t_j (as defined in Section 2.2), $l_{h,i}$ is the number of concepts of C_h found in d_i, and $tf \bullet idf()$ is a standard measure for computing the relevance of a term t_j in a document d_i of a collection D. Therefore each y_{ih} in v_i measures the overlap of d_i with the topic $C_h \in \mathcal{C}$.

We finally define a vector I_{g_i} which is the *centroid* of all publications vectors of g_i. The Content Based-Social Network is then modelled through an undirected graph with:

- the nodes representing the groups g_i;
- the edges representing the similarity between nodes, measured by the *cosine* function [17]:

$$\cos - sim(g_i, g_j) = \cos(g_i, g_j) = \frac{I_{g_i} \bullet I_{g_j}}{|I_{g_i}||I_{g_j}|}$$

4.2 Social Network measures

To analyze the network, we selected the following network analysis measures[10] [22]:

- *Average Degree Centrality*:

$$ADC = 1/(N(N-1)) \sum_{i=1}^{N} \deg(i)$$

where *deg(i)* is the number of edges connected to a node *i* and *N* is the number of nodes in the network. It measures whether the network is weakly or strongly connected.

[8] Evaluation on datasets in different applications makes no sense, since k-means++ has been already evaluated in the literature. What matters here is to measure the added value of terminology extraction and lexical chains.

[9] The clustering tendency of concepts is measurable by computing the *entropy* of the related similarity vectors. Sparse distribution of values over the vector's dimensions indicate low clustering tendency.

[10] For the sake of space, only two SN measures were selected here, leaving a more extensive analysis to forthcoming publications.

- *Degree Centrality of a Vertex*:

$$DC(v) = \deg(v)$$

It measures the connectivity of each social actor.

- *Weighted Degree Centrality of a Vertex*:

$$DC_w(v) = \sum_{e \in P_v} w(e)$$

where P_v is the set of edges connected to the node *v*, and *w(e)* is the weight of the edge *e*. It measures the connectivity of each social actor by taking into account the edges weight.

With respect to our similarity network, we can consider the *ADC* as a global measure of how far the community members share their research interests (i.e. the cohesion of the research community). The *DC(v)* and $DC_w(v)$ measure the potential for collaboration of each community member: in the first case, by considering only the presence of common interests between nodes, in the second, by taking into account also the similarity values.

4.3 Experiments on a research network

We conducted a set of experiments in which we applied the above measures to different networks obtained from different sets of partners' publications.

First of all, we calculated the *ADC* on four networks, obtained by grouping the 1452 papers written by the community members into four incremental sets, each of which contains, respectively, the documents produced before the end of year 2003, year 2004, year 2005 and until the end of the project. Table 2 summarizes the obtained results.

Table 2. ADC evolution over project duration

	Documents	ADC
Set1	595	0.5797
Set2	859	0.6773
Set3	1127	0.7855
Set4	1452	0.8744

The *ADC* values (see table 2) show how the shared research interests constantly increased during the project. Moreover, it is interesting to note that the highest increment of the *ADC* over a single time period was reached at the end of the second year of the project, the one in which the preliminary results of the partners' joint activities were obtained.

In a second experiment, we evaluated the topic distribution between groups, in order to identify the most popular research themes. Figure 4 shows the frequency of the 150 topics over the 46 I_{g_i} centroids acquired respectively as a result of topic clustering and CB-SN construction. The topics 5 and 56 have the higher frequencies (43 and 40 respectively). Not surprisingly (see discussion in Section 2.2) topic 56 is one of the "*ontology*" clusters:

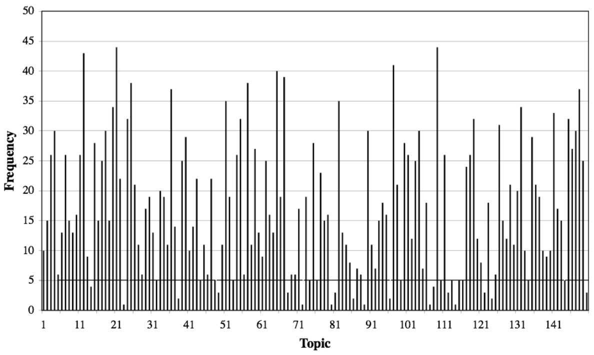


Figure 4. Topic popularity in the INTEROP research community.

```
{common_ontology, domain_ontology,
enterprise_ontology, informal_ontology,
ontology_alignment, ontology_analysis,
ontology_application, …}
```

since this sub-community is very well represented in INTEROP. Topic 5 deals with "*management*" concepts:

```
{architecture_management,
business_process_management, document_management,
enterprise_management, it_management, …}
```

The same type of analysis of figure 2 can be carried out at different time intervals, thus revealing shifts of interests and the emergence of new topics.

In a third experiment, all the 1452 documents were used and the $DC_w(v)$ for each node was calculated. Figure 5 is a view of the CB_SN interface in which the subnet of the global network, obtained by selecting only the edges with *cos-sim*$(g_i,g_j) \geq 0.6$, is shown. In the figure, the dimension of the nodes and the thickness of the edges are related, respectively, to the $DC_w(v)$ and the *cos-sim()* values. The biggest nodes represent the community members that have the highest potential in joint researches, whereas the thickest edges reveal the best potential partners. By clicking on the edges it is possible to visualize the topics involved in the similarity relation.

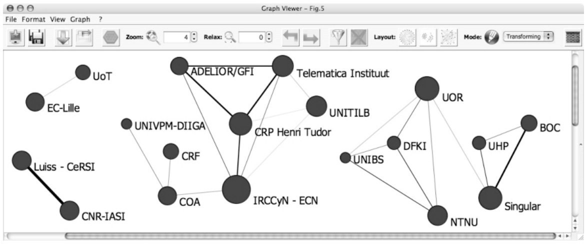
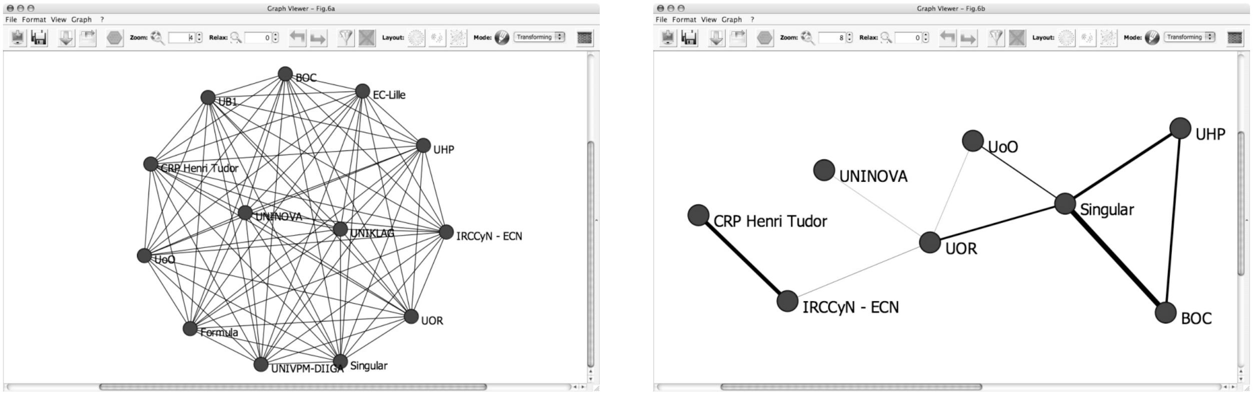


Figure 5. Graphic representation of DC_w in a subnet of strongly related nodes.



(a) (b)

Figure 6. A subnet of groups sharing the same research interest (a) and a highlight of the highest potential collaborations among them (b).

Another interesting analysis we carried out by using the interface was the selection of a topic and the visualization of all groups with research interests involving this topic. For example, starting from the global network of the previous experiment, we selected the subnet of all nodes sharing topic 15, associated to the cluster:

```
{composition_of_services, ontological_framework,
ontology-based_service}
```

Then we filtered the resulting graph by selecting the edges with *cos-sim*$(g_i,g_j)\geq 0.5$, to highlight the higher potential collaboration between groups involving this topic.

Figure 6a shows the complete graph obtained after the topic selection, and figure 6b the subnet of filtered potential collaborations where the thickest edges reveal the best potential links.

Finally, we use the SNA approach to give an insight into the real research partnerships among the groups. We modelled such relations through a “traditional” co-authorship network, where the edge between each pair of nodes has an associated weight that is the normalized number of papers co-authored by the members of the two groups. This value, $CP_{norm}(i,j)$, is defined as:

$$CP_{norm}(i,j) = \frac{CP(i,j)}{\min(P(i),P(j))}$$

where $CP(i,j)$ is the number of publication co-authored by the members of groups *i* and *j*, $P(j)$ is the number of publication of group *j* and $min(P(i),P(j))$ is the minimum value between $P(i)$ and $P(j)$. In this way $CP_{norm}(i,j)=1$ expresses the condition of maximum possible co-authorship between two groups (i.e. one group has all its publications co-authored by the other).

Figure 7 shows the network obtained by considering the 153 papers (over the 1452 written by the community members) co-authored by researchers belonging to different groups, in which the thickness of the edges is proportional to the $CP_{norm}(i,j)$ and the dimension of the nodes to the $DC(v)$. In the figure, it is possible to see, “at a glance”, the groups having the highest number of co-authorship relations (biggest nodes) with the others, and the pairs of groups having a high value of possible co-authorship (thickest edges), i.e. groups that have a strong collaboration in research activities. By comparing co-authorship and interest similarity data, the network analyser can identify those groups that have strong commonalities, but do not cooperate. This type of analysis is very useful e.g. in the diagnosis of research networks, like INTEROP, where one of the main objectives was to improve collaboration and result sharing among partners.

5. Concluding remarks

This paper addresses several novel aspects:

- It presents a social network analysis methodology highlighting the *content* of social relations, not

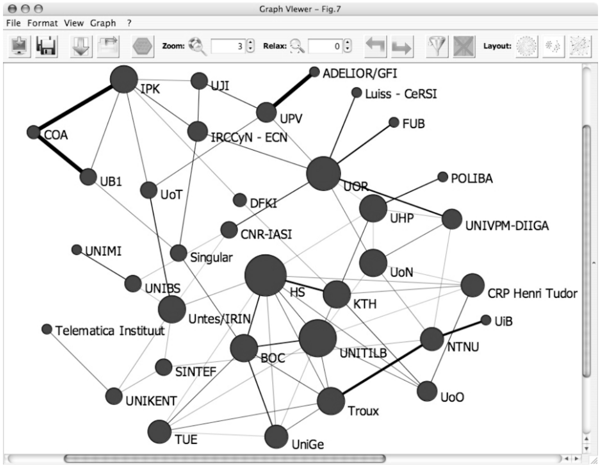


Figure 7. The co-authorship network.

merely the *number* of relations between actors. To the best of our knowledge, no similar methodology has been described in the social network literature.

- The communication content is modelled through clustering analysis, using: i) a tool to detect the *emergent semantics* of the social network domain; ii) a novel semantic measure of concept similarity, based on lexical chains of ontological and co-occurrence relations; iii) a high-performing variant of the k-means algorithm, k-means++.
- Finally, a visualization interface has been implemented to facilitate the study of the community by a social analyst.

Some relevant aspects and extensions of CB-SN are left to future work. Evaluation of topic clustering can be applied to document retrieval tasks, in order to assess the quality of the clusters obtained with our lexical chain method, in domains for which benchmarks are available. As far as a quantitative evaluation of the CB-SN model is concerned, this is per-se a new research topic, since the literature on social networks models presents only qualitative evaluations.

Furthermore, we aim to model a more complex social network, with two types of nodes: topics and researchers. By so doing, we can better analyze topic evolution across the time, a promising extension deferred to future publications.

6. References

[1] D. Arthur and S. Vassilivitski, “k-means++: The Advantages of Careful Seeding”, Proc. of the 18th ACM-SIAM Symp. on Discrete Algorithms, New Orleans, Louisiana,1027-1035, 2007.

[2] D. Bollegala, Y. Matsuo and M. Ishiuka, “Measuring semantic similarity between words using web search engines”, Proc. of the 16th international conference on World Wide Web, Banff, Alberta, Canada, 2007.

[3] U. Brandes, M. Gaertler and D. Wagner, “Engineering Graph Clustering: Models and Experimental Evaluation”, ACM Journal of Experimental Algorithmics, Vol. 12, 2007.

[4] Dhiraj, J., Gatica-Perez, D., “Discovering Groups of people in Google News”. In Proc. of HCM’06. Santa Barbara, CA, USA, 2006.

[5] W. A Gale, K. W. Church and D. Yarowski, “One sense per discourse” in Proc. of the Workshop on Speech and Natural Language, Harriman, NY, 233-237, 1992.

[6] K. Hammouda and M. Kamel, “Efficient Phrase-Based Document Indexing for Web Document Clustering”, IEEE Tr. On Knowledge and Data Engineering, IEEE Transactions on Knowledge and Data Engineering, Vol.16, N. 10, 1279-1296, (2004).

[7] G. Hirst and A. Budanitsky, “Lexical Chains and Semantic Distance”, EUROLAN-2001, Romania, 2001.

[8] A. Jain, K. M. Murty and P. Flynn, “Data Clustering: A Review”, ACM Computing Surveys, Vol. 31, N. 3, 264-323, (1999).

[9] M. Jamali and H. Abolhhassani, “Different Aspects of Social network Analysis”, Proc. of the 2006 IEEE-WIC-ACM Int. Conf. on Web Intelligence, 2006.

[10] S. Kang, “Keyword-based document clustering”, Proc. of the 6th int. Workshop on Information Retrieval with Asian Languages, Vol. 11, 132-137, Japan, 2003.

[11] F. Kovacs, C. Legany and A. Babos “Cluster Validity Measurement Techniques” 6th Int. Symposium of Hungarian Researchers on Computational Intelligence, November 18-19, Budapest, Hungary, 2005.

[12] A. McCallum, A. Corrada-Emmanuel, X. Wang, “Topic and Role Discovery in Social Networks”. In Proc. of Int. Joint Conf. on Artificial Intelligence (IJCAI), pp. 786-791. Edinburgh, 2005.

[13] G. Nenadic, S. Rice, I. Spasic, S. Ananiadou and B. Stapley, “Selecting Text Features for Gene Name Classification: from Documents to Terms”, Proc. of the ACL 2003 Workshop on NLP in Biomedicine, Vol. 13, 121-128, Sapporo, Japan, 2003.

[14] M. E. J. Newman, “The structure and function of complex networks”, SIAM Review 45, 167-256, (2003).

[15] T. Pedersen, S. V. Pakhomov, S. Patwardhan, C. G. Chute, “Measures of semantic similarity and relatedness in the biomedical domain”, Journal of Biomedical Informatics, Volume 40 Issue 3, 288-299, (2007).

[16] P. Resnik, “Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language”, Journal of AI Research, Vol. 11, 95-130, (1999).

[17] G. Salton and M. McGill, “An Introduction to Modern Information Retrieval”, McGraw-Hill, New York, 1983.

[18] F. Sclano and P. Velardi, “TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities”, Proc. of 9th Conf. on Terminology and Artificial Intelligence (TIA 2007), Sophia Antinopolis, 2007.

[19] P. Tan, M. Steinbach and V. Kumar, “Cluster Analysis: basic concepts and algorithms” in Introduction to Data Mining, Addison-Wensley, 2006.

[20] E. Terra and C. L. Clarke, “Frequency estimates for statistical word similarity measures”, in Proc. of the 2003 Conf. of the North American Chapter of the ACL on HLT (NAACL '03), Morristown, NJ, 165-172, 2003.

[21] P. Velardi, A. Cucchiarelli and M. Petit, “A Taxonomy learning Method and its Application to Characterize a Scientific Web Community “, IEEE Transaction on Data and Knowledge Engineering (TDKE), Vol. 19, N. 2, 180-191, (2007).

[22] S. Wasserman and K. Faust, “Social Network Analysis: Methods and Applications”, Cambridge University Press, 1994.

[23] J. Weeds and D. Weir, “Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity”, Computational Linguistics, Vol. 31, N. 4, 439-475, 2006.