2009 Ninth International Conference on Intelligent Systems Design and Applications

Equalizing the Structures of Web Communities in Ontology Development Tools

Francesca Arcelli Fontana, Ferrante Raffaele Formato, Remo Pareschi
University of Milano Bicocca, University of Sannio, University of Molise
arcelli@disco.unimib.it, ferrante@unisannio.it, rpareschi@ngponline.com

Abstract

In this paper we face some relevant issues on the relations between web communities and ontologies. We build an operator that constructs a weak Web Community, according to the definition given in [16],starting from a seed of web sites. The necessity of such an operator is derived from a problem arisen in the model developed in [3], in which some relevant concepts in automotive oriented ontology were not given a corresponding Web community. This fact –if not considered- can bring automatic ontology development ([9,18]) to some non-correct results.
In this work we define and analyze a new operator, called Com, with the tools furnished by the method of parametrization ([8,15]) and we find that, given a seed S and the induced graph I(S), the community generated by our operator is monotonic with respect to clustering and is denser than the original graph I(S).

1. Introduction

At present, ontologies are built manually by a panel of experts and the tools for assisting automatically the construction of new ontologies are based on text analysis. In [3, 4] we described a tool for building ontologies that locally reflects Web communities and we started an experimental ontology describing the field of automotive. The experience with our prototype Gelsomino [11] revealed that some relevant concepts in the field of automotive - such as "SCR-based ecologically sustainable engines [14]" - were not tokenized enough. On the contrary, sites about "restaurants" and "regional tourism" were hubs of many automotive communities. Of course, if we extract these concepts and put them into an automotive oriented ontology, then such ontology is not longer reliable. Therefore, we need instruments that equalize the structures of a web community. We have been studying the problem of ontological alignment ([9]) and we found that ontological alignment is hard to do without mapping communities to concepts.

FaceBook and MySpace have been developing some simple tools for balancing the structure of a community, based on the "small-world" assumption ([1]): *If a is a friend of c and b is a friend of c, then a and b could be a friend of d, too*.
But this is causing a lot of problems in social networks. In fact, if *a* and b are males looking for a mate and c is female, then a and b will hardly get along well. On the contrary, the dynamics of èlite network, *a small world.Net,* takes in consideration only the principle of network synchronization [2]–or co-evolution [3] - and they say that their Web Community is highly robust and top-level quality.

Network parameterisation has been intensively studied in [8, 15]. In this paper, a set of parameters is drawn and through them we propose a characterization of "good networks"; the parameters are the following:

- Clustering
- Diameter
- Average degree
- Degree distribution
- Spectrum

Network parameterisation is a diffused area of research. In [7] a feature vector of parameters is associated to a complex network. By associating a feature vector of parameters in R^n –seen as a space state- to a network, we can model the evolution of a network by a path in a space state. These results arise the following problem in our model: given a concept *c* with little tokenization but significant information, is there a network operator that equalizes the associated set of web sites $S(c)$ by extending it into a "good network"?
Good networks are networks without massive connected components that trends to monopolize the whole network. This passage of state is called "percolation" and it is a quiescent state of a network. In percolation a complex network is divide into a few

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.223

massive components. Some authors assert that the quality of a network is given by the percolation threshold ([17]).
In the present paper we build such an operator and we check its validity by network parameterisation methods. In this way we can draw some considerations on the definition and classification of a network as a "good" network.

The paper is organized through the following Sections: In Section 2 we introduce complex networks and we definer *scale free networks* and *small world networks*, in Section 3 we introduce the parametrization of complex networks, in Section 4 we define a graph operator for building web communities and it is described how scale-free networks are transformed into scale-free communities, in Section 5 we introduce the network parametrization of the community operator defined in section 4. Finally in Section 6 we conclude our work outlining some future developments.

2. Complex Networks

Although there is a general agreement on considering as "complex" for example the networks of airline routes, the networks of power distribution and the networks of social relations, at present there is not a general accepted definition of "complex network", probably because such definition is "complex" itself.
According to Erdös and Rényi ([10]), in a random graph with N nodes and connection probability p, the probability $P(k_i=k)$ that node i has degree k follows a binomial distribution. In fact we have:

$$P(k_i = k) = C_{N-1}^{k} p^k (1-p)^{N-1-k}.$$

In this work we adopt the following definition: a network is *complex* when the distribution of nodes and edges does not follow the model of Erdös and Rényi.

A network is *scale free* when the degree of distribution of its nodes does not depend from the size of the network. In particular, Barabasi and Réka [5] showed that in scale-free networks, the following holds:

$$P(k_i = k) = \frac{1}{k^{\gamma}}.$$

Where i is the node and $P(k_i=k)$ is the probability that a new arc is added to a node with degree k_i. This means that a real network is not a random graph. Rather, the percentage of nodes with, for example, 8 arcs is:

$$p(8) = \frac{1}{8^{2.1}} \approx 0.15.$$

Where γ is a real number in the interval [1,2].

At present there is an ever growing tendency to search for network parametrization. [7, 13] and the network is called *complex* if the degree distribution is not poissonian. Thanks to the works of Watts and Strogatz [19] and Barabasi and Reka [5], a taxonomy of complex networks has being compiled.

In particular, two categories of networks have emerged:

-) scale free networks.

A scale free network is organized independently on the number of its nodes and contains few hubs –nodes with a great number of edges-. Examples of scale-free networks are air routes, power distribution and Internet.

-) small world networks.

In small-world networks any pair of nodes is not far. In [19] a small-world network is described as a regular lattice combined with a random graph. Therefore, in small world networks nodes are highly clustered. There are edges –called *weak ties*- that are distributed randomly throughout the network. The importance of weak ties has been studied in [12]. Weak ties are useful for example to find a job. Typical examples of small world-networks are social networks and World Wide Web [1].

3. Complex Networks and their parametrization

We now describe a set of parameters that we used to classify a complex network.

3.1. Degree

The *degree* of a node of a network is the total number of its connections. The in-degree (resp. outdegree) is the number of incoming (resp. outcoming) edges. By indicating with k, k_i and k_o the degree, in-degree and out-degree of a vertex, we have:

$$k = k_i + k_o$$

We indicate with $P(k)$, $P(k_i)$ and $P(k_o)$ the distribution of the degrees, in-degrees and out-degrees, respectively.

3.2 Shortest path

Let G be a graph. Given a pair of nodes μ and ν, we call *geodesic distance* $d_g(\mu,\nu)$ the length of the shortest-path in G between μ and ν.
Observe that if G is a directed graph then d_g is not a distance.

We now formalize the concept of small-world network.

Let Ω_n be the set of all the networks with n vertices and N as set of vertices. Let $\omega \in \Omega_n$. We call $\pi(\mu,\nu,\omega)$ the distribution of the shortest paths over Ω_n. Then we define the *average shortest path in* ω as

$$< l, \omega >= \sum_{\mu,\nu \in N_\omega} \pi(\mu,\nu,\omega) d_g(\mu,\nu) \quad .$$

Finally,

$$< l >= \sum_{\omega \in \Omega_n} < l, \omega >$$

$<l>$ is also called the diameter of the network. It can be shown that:

$$< l > \approx \frac{\ln n}{\ln z} .$$

Where n is the cardinality of the nodes of the network and z is the average number of nearest neighbours of a vertex. By setting:

$$\sigma = \frac{\ln n}{\ln z}$$

we capture the property of *small world* , as observed in [19].

3.3 Clustering

We recall that a graph $G = (V,E)$ can be partitioned into a family of set of nodes $\{V_i\}_{i \in V}$. It is obvious that, in a network with n nodes, the number of possible connections is $\frac{n(n-1)}{2}$. Let σ be the fraction of actual connections in the networks. The ratio between the number σ and the clique:

$$C_\lambda = \frac{2\sigma}{n(n-1)}$$

is the clustering coefficient with respect to node i. Then the clustering coefficient of a network is:

$$\sum_{\nu \in N} \frac{C_\nu}{n} .$$

In a random graph, Watts and Strogatz [19] define a statistic parameter.

$$< C >= \sum_{i \in N} P(i) C_i .$$

Clustering is an important parameter when we want to search for Web communities. In fact, it measures the average number of clusters in the graph. In fact, when $<C>$ is not high, it is unlikely that communities can be found in the graph.

3.4 Edge degree distribution

Another important parameter of a network is the edge distribution. In a random graph this distribution is poissonian, i.e.

$$P(k_i = k) = C_{N-1}^k p^k (1-p)^{N-1-k} .$$

Barabasi and Rèka ([5]) have shown that in many networks such distribution is not uniform, but is a function with a tail exponentially decreasing.

$$P(k_i = k) = \frac{1}{k^i} .$$

Our aim is setting up a model that equalizes ontologies with complex networks, whose evolution is controlled by these parameters.

Network parameters are used in the construction of meta-ontology [4]. We must associate to each concept a union of graphs.

3.4 Spectrum

The spectrum of a network G is the set of eigenvalues of the matrix associated to G. In particular, as shown by Kleinberg in [13], the principal eigenvalue individuates a pair of vectors **h** and **a**. Probably, the remaining eigenvalues characterize the clustering of

the graphs. In particular, given a secondary eigenvalue λ , for any eigenvector $\mathbf{v}$ of λ there exist a pair of integers i and j such that $v_k\ k \geq i$ and $v_k\ \ k \leq j$ are a cluster in G ([13]).

4. A graph operator for building Web communities.

Web communities are studied in social sciences, biology and computer science. The automatic extraction of community out of a complex network requires a formal definition of community. At present there is not a general agreement and several definitions of Web community have been proposed.
We try to report them and recall that an *N-clique* is a subgraph G' such that any pair of nodes in G' is linked by a path of length not greater than n.

Definition 4.*1* *Let n be an integer. An n-web community is a n-completely connected subgraph, or n-clique.*

In Figure 1 and Figure 2 several examples of Web communities are depicted. Figure 1 (a) is a 1-clique or 1-Web community. Figure 1-(b) is a 2-clicque or 2-Web community, Figure 2-(c) is a strong Web community and Figure 2-(d) is a weak Web community.

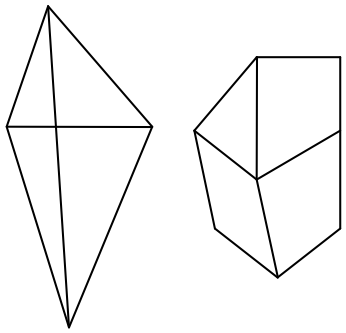
Definition 4.2 *Let G be a graph. A strong web community is a subgraph G'=(N',E') of G such that, for any i∈N', the internal links of G' are greater then the external links of G'.*

Finally, the following is an extension of definition 4.2.

Definition 4.3 *Let G be a graph. A weak web community is a subgraph G'=(N',E') of G such that the sum of internal links in G' is greater then the sum of external links in G'.*

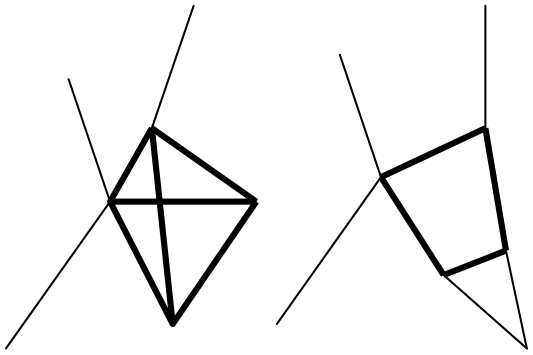
We now define an operator that, given a set of nodes S and the induced graph $I(S)$, takes $I(S)$ as input and yields a set of smallest web communities containing $I(S)$.

Definition 4.4 *Given a graph G = (V,E), we define the operator Com and we set Com(G) as follows*:
Until G is a strong web Community
Do
Begin
For any node v, let Out(n) be the number of outlinks of n.
IF Out(n) > In(n) **then**
Add Out(n) –In(n) edges to the n+1….n+|Out(n)-In(n)| outside G
Close (N,E)
End



(a) (b)

Figure 1. A Taxonomy of Web Communities



(c) (d)

Figure 2. A Taxonomy of Web Communities

The operator *Com* transforms a graph G into a weak community by adding to each node the number of internal edges in G necessaries to outnumber the external edges in G.

Proposition 4.1

i) Com(G) is a Web community
ii) If G' is a Web community and G'⊇G then G' ⊇ Com(G)

Proof: Immediate.

We immediately observe that the distribution degree is not an invariant of operator Com. In other words, the degree distribution of the graph G, that we indicate with P_G, is different from the degree distribution of Com(G), that we indicate with $P_{Com(G)}$. To prevent this, we define a new –non-deterministic –operator Com'.

Definition 4.5. Given a graph $G = (V,E)$, we set Com'(G) as follows:

Until G is a web Community
Do
Begin
For any node v , let Out(n) be the number of outlinks of n.
IF Out(n) > In(n) **then**
Chosed Out(n) –In(n) n+1….n+|Out(n)-In(n)| according to P_G
Add Out(n) –In(n) edges to the n+1….n+|Out(n)-In(n)| outside G
Close (N,E)
End

The random choosing can be made as follows:

Given P_G, generate a pseudo-random bit a
If $a = 1$ **then** add the node to Com_G.
If $a = 0$ **then** do not add the node.

The random choosing is made in order to preserve the distribution degree of the graph. By so doing, scale-free networks are transformed into scale-free communities. The ontology affects the network because the network grows around the concepts of the ontology

5. Network Parametrization of the Community Operator

We now analyze the operator *Com* with respect to the parameters of section 3. By so doing, we prove that the graph of the community is denser then the induced graph of the seed *G*.

Proposition 5.1 Let *G* be a graph. Then

$$< C(Com(G)) > \geq < C(G) >$$

Proof.

By definition, we have that:

$$< C >= \sum_{i \in N} P(i)C_i .$$

By construction,

$$< Com'(G) >= \sum_{i \in Com'(N)} P'(i)C_i$$

Since $Com'(N) \supseteq N$ and $P(i) = P'(i)$ we have that:

$$\sum_{i \in Com'(N)} P(i)C_i \geq \sum_{i \in N} P(i)C_i$$

The thesis follows.

Proposition 5.1 shows that the graph yield by Com is denser then the input graph. Clustering preservation is a sufficient condition for the following property:

Proposition 5.2 Let *W* be a Web community and $W \subseteq I(S)$ Then

$$<C(W)> \leq C(Com'(I(S)))$$

Proof. Immediate.

Proposition 5.2 says that when operator Com' is applied to a set S that is mostly rarefied but contains some Web community *W*, it yields a community *I*(*S*) denser than *W*. This is shown in Figure 2.

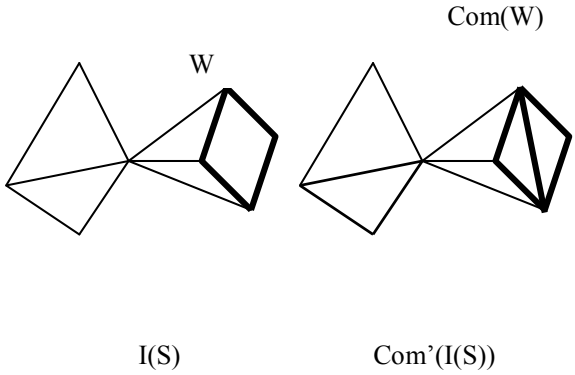


Figure 3. Clustering Effects.

Figure 3 shows the clustering effects of the operator Com' on Web communities: we can see that community Com(W) is still a community but is denser then W.
Network parametrization shows that both operators Com and Com' can be used to equalized scanty seeds.

6. Conclusions

We have introduced a methodology –complex network parametrization- that studies the evolution of complex networks and we have constructed an operator on

graphs. The purpose of our operator is equalizing meta-ontologies in our model [4]. In particular, operator Com transforms a scattered group of web sites -seed- associated to a (very) important concept of an ontology into a robust and dynamic complex network.
We must now investigate the behaviour of the seed with respect to other graph operators, according to the relaxed definition of Web Community.
We recall that our main goal is developing a computational model of network evolution controlled by network parametrization and interfaceable with ontologies.
Effective tools for ontologies construction must operate on "good networks". Therefore a necessary condition for building these tools is being able to discern "good" networks. We know that a scattered network is not a good network. Therefore, clustering is a parameter that characterizes a good network. Many scale-free networks have proven to be resistant to attacks and therefore we need an operational characterization of this notion. For example, in [17] the distribution degree of a scale-free network has been related to the percolation threshold, but for a finer grasping of clustering, we need definitions based on network entropy.

References

[1] Adamic L. A. The Small World Web, in *Lecture Notes on Computer Science* , 1696, pp. 443-452, 1999.

[2] A. Arenas, A.D. Guilera, J. Kurts, Y. Moreno and C. Zhou, Syncronization in Complex Networks, *Physics Reports,* 469, pp. 93-153, 2008..

[3] F.Arcelli, F.Formato and R.Pareschi, Ontology Engineering: Co-evolution of Complex Networks with Ontologies*, Proceedings of the First Workshop on Ontologies for e-Technologies(OET), May 2009* .

[4] F.Arcelli, F.Formato and R.Pareschi. *Reflecting Ontologies into web* Communi*ties, IEEE Conference IAWTIC*, Vienna, December 2008.

[5] A.L.Barabasi and A. Réka, *Emergence of Scaling in Random Networks*, *Science,n.* 286, pp. 509-512, 1986,

[6] T. Berners Lee, N. Shadboldt and W. Hall, *The Semantic Web Revisited*, *IEEE Intelligent Systems*, 21 (3), pp. 96-101, 2006.

[7] Brin, S., Page, L.: *The Anatomy of a Large-Scale HypertextualWeb Search Engine*
http://google.stanford.edu/long321.htm

[8] L.F. Da Costa F.A. Rodrigues G. Travieso and P.R. Villar Boas Characterization of Complex Networks:A Survey of measurements. *Advances in Physics* , 56, January 2007.

[9] C. De Maio, G. Fenza, V. Loia, S. Senatore, Ontology-based knowledge structuring: an application on RSS Feeds, *2nd International Conference on Human System Interaction*, Catania, Italy, May 21-23, 2009.

[10] P. Erdös and A. Rényi *Random Graphs,* Publicationes Matematicae Inst. Hung. Acad. Sc, 5, 1960.

[11] Gelsomino: Gelsomino is available for download at www.essere.disco.unimib.it/ on request.

[12] M. Granovetter The Strength of weak ties, *American Journal of Sociology*, Vol. 78, May 1973.

[13] J. Kleinberg., Authoritative Sources in an Hyperlinked Environment, *Nature 406*, 2000.

[14] Iveco Press Release http://www.iveco.com/en-us/PressRoom/PressRelease/Pages/1073758784.aspx.

[15] M. Mihail, Gkantisidis, A. Sabeni and Zumara, *On the semantic of Internet topologies* , Gatech Technical Report, 2009.

[16] F. Radicchi, C. Castellano, F. Cecconi V. Loreto and D. Parisi, Defining and Identifying Communities in a Network, *Proceedings of National Academy of Sciences, 1* 01 pp.2658-2663 ,2004.

[17] P.Satorras R.Vespignani, Epidemic Spreading in Scale-free network, *Phisical Review Letters*, 86-14, pp.3200-3203, 2001.

[18] Q. T. Tho, S. C. Hui, A. C. M. Fong, and T. H. Cao, Automatic Fuzzy Ontology generation for Semantic Web, *IEEE Transactions on Knowledge and Data Engineering,* Vol. 18(6), pp. 842, 2006.

[19] D.Watts, S.Strogatz, Collective dynamics of 'small world' networks. *Nature* 393, 1998.