Clustering Enterprise Networks by Patent Analysis

Fulvio D´Antonio[1], Simone Orsini[1], Alessandro Cucchiarelli[1], Paola Velardi[2]

[1] Polytechnic University of Marche, Italy, email: cucchiarelli, dantonio, orsini@diiga.univpm.it
[2] Sapienza Università di Roma, Italy, email: dantonio,velardi@di.uniroma1.it

Abstract. The analysis of networks of enterprises can lead to some important insights concerning strategic aspects that can drive the decision making process of different players: business analysts, entrepreneurs, public administrators. In this paper we present the current development status of an integrated methodology to automatically extract enterprise networks from public textual data and analyzing them. We show an application to the enterprises operating in the Italian region of Marche.

Keywords: Social Network Analysis, Natural Language Processing, Clustering

1 Introduction

Networks of Enterprises [2] are a special kind of social networks in which the nodes represent enterprises and the links indicate some form of relationship among them.

The relationships that have been traditionally represented through links are business collaborations, enterprise similarity, mutual exchange of capitals, information flows, or hierarchical relationships like the ones representing supply chains or enterprises aggregation into districts.

Social Network Analysis [6] defines a number of measures and techniques that can be used for the evaluation and analysis of enterprise networks. Such measures, if examined by a business analyst, an entrepreneur or a public administrator can lead to some important insights concerning some strategic aspects of the network.

We describe here few scenarios in which the analysis can be conveniently applied:

- Domain analysis

The analyst inspects the network in order to understand which are the main productive sectors, the groups of similar enterprises, the relative strengths of such groups and their inter-relationships.

- Determining competitors

Mining non-cooperating similar enterprises which may be potential competitors in a given productive sector. There is either high or low level of competition? There is a potential for market penetration of my enterprise?

- Partnership discovery

Individuating similar or complementary enterprises aimed at establishing business/productive co-operations.

- Funds allocation

Analysis of productive trends and gaps, and setup of regional/national funding schemes.

But where the data about Networks of Enterprises come from?

The usual scenario is that the graph structure of the network is not explicitly available but has to be "distilled" from a dataset *D*, i.e., one has to infer the network structure starting from such data by applying some processing steps.

Let's examine, as an example, the case of networks whose (weighted) links represent the degree of "similarity" between the nodes. We have two possibilities:

1. We can submit questionnaires to the actors involved asking them to estimate their similarity with, let's say, one hundred of other enterprises. The similarity value could be a real number in the range [0,1], a set of symbols (sequence of stars, for example: * little, ** medium , *** high or no stars for no similarity) or similar representations.
2. If we have some textual data available, e.g. papers, websites, product manuals etc. we can use some form of natural language processing and information retrieval metrics to (semi)-automatically estimate the similarity.

The first approach is expensive, exposed to questionnaire's compiler subjectivity and implies a series of practical issues: distribution of the questionnaires, commitment to the questionnaire compilation in a given time and collection of the results.

The second approach enjoys the benefits of the general wealth of publicly available data and of automatic processing; everyone can search the web and obtain a great number of information (mainly textual) about the enterprises under examination. The drawbacks of this approach rely in the generally worse performance of natural language processing systems with respect to humans. Humans seems to be better in performing tasks like word-sense disambiguation, contextualizing judgement and understanding the textual information.

Hybrid approaches are also commonly adopted: an automatic NLP system interact from time to time with humans that take decisions about some harsh points.

Let's consider an enterprise interested in finding potential partners among the enterprises in a given geographical area, that, in turn, requires to find partners with similar interest. Even in small areas the enterprises, generally mostly SMEs, (Small-Medium Enterprises) can easily be in the order of several hundreds. If we decide to assign such task to a person we could apply the following strategy: we give him/her a list of some hundreds of enterprise names and some thousands of documents and related websites and we ask him/here to read the documents and surf the websites to extract key information about the business/productive sector of the enterprise in order to estimate from such information the degree of similarity and potential collaboration. This task is clearly not feasible for a human. A valid support can come from a carefully designed NLP system that can be supervised by the user and occasionally corrected by him/her (e.g. eliminating non-relevant keywords in a particular domain, individuating uncaught spelling variation, etc).

2 Patent and Enterprise Networks

In this section we describe how we have distilled Networks of Enterprises starting from textual data publicly available about patents deposited by European enterprises.

The European Patent Office (EPO)[1] provides a uniform application procedure for individual inventors and companies seeking patent protection in up to 40 European countries. It is the executive arm of the European Patent Organisation and is supervised by the Administrative Council. Through its web-site and exposed web-services it is possible to access to information about European patents that have been registered; the information include, among the other things, the date of presentation, the applicant name and mission, the address of the applicant and the textual description of the patent.

The patents presented by an enterprise is a good indicator of the business sector in which the enterprise operates. Therefore through the EPO database we can gather textual data about the business/industrial sector of the enterprises in a given geographical location and we can use such data to extract similarity networks. The methodology we use is summarized in the following steps and it is similar to the ones used in [4,5]:

1. Gather patents registered by enterprises located in a given geographical area (a city, a region, a country, …);
2. Pre-process textual data to extract raw text;
3. Process raw text with a part-of-speech tagger;
4. Extract candidate annotating terms using a set of part-of-speech patterns [3];
5. Rank candidates, possibly filter them choosing a threshold [3];
6. Output a set of weighted vectors V of annotating terms for each documents;
7. Group the vectors by enterprise (that presented the patent applications) and construct a centroid (i.e. a mean vector) with such groups. This centroid roughly represents the business sector of the enterprise.
8. Build a graph computing a similarity function [1] for each pair of centroids.

2.1 Clustering

Data Clustering [8], originally conceived in the data mining field, is a very active research domain aiming at developing methods for dividing a set of data-points into subsets (called clusters) so that points in the same cluster are similar in some sense. We can use clustering techniques on our Enterprise Networks in order to discover potentially interesting networks patterns and to filter noisy phenomena.

[1] http://www.epo.org/

One of the main drawbacks of clustering is the substantial lack of possibility of validating results except for very special cases, e.g. when the distribution of data is known (like a multivariate Gaussian) or we have access to other forms of ground truth. In literature clustering validation is approached using internal and external validity criteria: the external criteria rely on comparison with available ground truth while the internal ones are constituted by metrics that estimate the internal coherence of a cluster (inter-cluster similarity) and its substantial dissimilarity from other clusters (intra-cluster dissimilarity). According to [7], each clustering technique should be evaluated in the context of a micro-economic setting, i.e. in maximizing an objective function.

We relax as much as possible the notion of clustering: given a set *A*, a clustering *C* is a set of subsets of *A*, i.e. $C \subseteq \mathrm{P}(A)$ where P(A) is the power set of A. A *crisp clustering* is a clustering with pairwise disjoint clusters and a *partitive clustering* is when the union of clusters is A ($\bigcup_{C_i \in C} C_i = A$).

Most of the clustering techniques developed concentrate on producing *partitive crisp clusterings*.

2.2 Graph clustering by mean of components density maximization

In this paper we use a very simple algorithm for graph clustering. Given a graph *G=(V,E)* in which *V* is a set of vertices and *E* is a set of weighted edges *(x,y,w)* with *x,y* in *V* e *w* in *[0,1]*, we order the edges in *E* with respect to the weights obtaining the sequence $e_1,\ldots,e_{|E|}$. We then construct the sequence of graphs $GS=G_0,\ldots G_{|E|}$ in which $G_i=(V,\{e_1,\ldots,e_i\},$ i.e. the *i*-eth graph is the graph containing the top-i weighted edges. The clusters are the connected components of each graph and each graph contains all the others following in the sequence so that, therefore, we have a hierarchical clustering.

To choose a representative of this sequence we maximize the function scoring the *mean components density*: for a graph we compute the density of each connected component, we sum them and we divide by the number of components. The (weighted) density of a connected graph is:

$$d(G) = \frac{\sum_{(x,y,w) \in E_G} w}{\binom{|V|}{2}}$$

The *mean components density* is:

$$meand(G) = \frac{\sum_{C \in Components(G)} d(C)}{|Components(G)|}$$

And finally, we can choose the preferred clustering G_{pref} by maximizing *meand*:

$$G_{pref} = \underset{Gi \in GS}{\arg\max}\, meand(G_i)\,.$$

3 Applications

In figure we show a detail of the graph obtained by applying the described method to the enterprises operating in the Italian region of Marche that registered European patents. The graph has been clustered according to the algorithm in section 2.2.

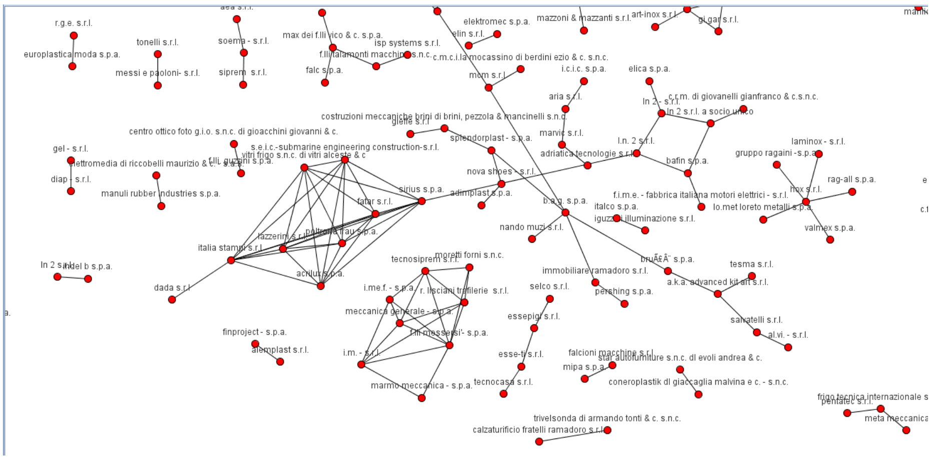

Fig. 1. The Network Of Enterprises of Region Marche (detail)

In the figure we can visually locate a very dense cluster in the middle-left; unfortunately an in deep analysis of this clusters reveals that it is consisting of all enterprises that deposited patents in German language. At the beginning of the experimentation we didn't notice that some patents descriptions are not written in English language. This noisy phenomenon, anyway, emerged because of clustering and we suggest that this can become one important use of clustering techniques: locating "spam" clusters in order to eliminate them and iteratively refine the process.

In the rest of the picture we notice a high degree of fragmentation: several very small groups (2 or 3 elements) and rare bigger groups.

We report here some examples of clusters:

- Moretti forni S.p.a
- Defendi Italy S.r.l
- Officine Meccaniche Defendi S.r.l
- S.o.m.i press

In which the similarity links depend mainly on the terms: *gas, flame, burner, cooking*. We can suppose this is a cluster consisting of cooking-furniture enterprises.

Another cluster is constituted by:

- Best S.p.a
- Gitronica S.r.l
- Intec-s.r.l

depending on the terms *phone, microphone, voice, electronic component*.

In general is very difficult to evaluate the quality of the produced clusters and we performed only a qualitative analysis.

A high level of fragmentation is, indeed, a problem. The utility of clustering in general is to reduce the dimension of problems: if the number of clusters is comparable with the number of elements we haven't performed any reduction at all and the clustering is useless. As we performed just an initial experimentation we are not able to say if the fragmentation observed is a real phenomenon in the application domain or can be reduced by refining the techniques used in the various steps of the process.

Therefore, in the future, we plan to work on the following points:

- The NLP analysis tools and techniques we adopt are powerful enough to put in light important similarities/differences in the domain studied?
- The data used are enough complete/noise-free/etc? If not, how can we perform data cleaning and gather additional data?
- The clustering method proposed is comparable with respect to state-of-the-art methods?

4 References

1. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Modern Information Retrieval, Addison Wesley, 1st edn., May 1999.

2. T. Elfring and W. Hulsink, 'Networking by entrepreneurs: Patterns of tie-formation in emerging organizations', Organization Studies, 28(12), 1849–1872, (2007).

3. Francesco Sclano and Paola Velardi, 'Termextractor: a web application to learn the shared terminology of emergent web communities', in Proceedings of the 3rd International Conference on Interoperability for Enterprise Software and Applications (I-ESA 2007), Funchal, Portugal, (2007).

4. Paola Velardi, Alessandro Cucchiarelli, and Fulvio D'Antonio, 'Monitoring the status of a research community through a knowledge map', Web Intelli. and Agent Sys., 6(3), 273–294, (2008).

5. Paola Velardi, Roberto Navigli, Alessandro Cucchiarelli, and Fulvio D'Antonio, 'A new content-based model for social network analysis', in ICSC '08: Proceedings of the 2008 IEEE International Conference on Semantic Computing, pp. 18–25, Washington, DC, USA, (2008). IEEE Computer Society.

6. Stanley Wasserman, Katherine Faust, and Dawn Iacobucci, Social Network Analysis : Methods and Applications (Structural Analysis in the Social Sciences), Cambridge University Press, November 1994.

7. J. Kleinberg, C. Papadimitriou, P. Raghavan. A micro-economic view of data mining. Data Mining and Knowledge Discovery, 2(4), 1998.

8. Ian H. Witten, Eibe Frank , Data Mining: Practical Machine Learning Tools and Techniques (Second Edition), Morgan Kaufmann June 2005