Topic connections and clustering in text mining: an analysis of the JADT network

Domenica Fioredistella Iezzi

Department of Philosophical Studies -Tor Vergata University of Rome

Riassunto

Nei convegni scientifici, gli studiosi discutono lo stato di avanzamento dei loro lavori e instaurano nuove collaborazioni. Le conferenze, quindi, costituiscono un luogo privilegiato dove avviare dei legami con altri studiosi e promuovere nuove attività di ricerca. La collaborazione tra ricercatori rappresenta un prototipo di reticolo sociale. Nel presente lavoro, sono stati esaminati i paper presentati nelle dieci edizioni del Convegno JADT dal 1990 al 2010. L'obiettivo della ricerca è stato di individuare i metodi più utilizzati delle diverse edizioni. In particolare, sono state evidenziate le strutture di collaborazione a livello macro (la rete globale che si è creata nelle diverse edizioni) e a livello micro (gli attori centrali e i temi più rilevanti). A questo scopo sono state utilizzate le tecniche di Social Network Analysis per misurare la centralità di alcuni nodi e classificare gli argomenti trattati dagli studiosi.

Abstract

In academic conferences, researchers usually discuss their work and clasp scientific collaborations. Conferences provide an important link for the exchange of information among scholars. Collaborations among scientists represent a prototype of social network. We present the macro (the whole network) and micro (both actor and topic centred) structure of collaboration in the last 20 years in International Conference JADT. We use centrality measures to evaluate roles of scholars and their topics and classify research topics.

Key-words: centrality measures, cluster analysis, Social Network Analysis, text mining

1. Introduction

Collaboration among scholars is a key mechanism of knowledge flows in research activities. It represents a complex phenomenon, which should be studied from several perspectives: individual scientists, research institutions, national and international research polices. Several studies have shown that scientific productivity depends, among other things, on scientists' attitude towards collaboration in research (Lee and Bozeman, 2005; van Rijnsoever et al., 2008). Text mining is a new research area that requires expertise in Information Technology, Linguistics and Statistics (Bolasco, 2005), because it tries to solve problems by using techniques from data mining, machine learning, natural language processing (NLP), information retrieval (IR), and knowledge management (Feldman and Sanger, 2007).

The aim of this paper is to propose a method to analyse the macro (the whole network) and micro (both actor and topic centred) structure of collaboration in the last 20 years in International Conference on statistical analysis of textual data conferences (JADT). We want to discover patterns of collaboration in a closed network of scholars and to find links in writing articles. Cluster Analysis encompasses many different techniques for discovering structure within data. Gordon (1999: 1) remembers that "the subject of classification is concerned with the investigation of the relationships within a set of "objects" in order to establish whether or not the data can validly be summarized by a small number of classes (or clusters) of similar objects». All clustering problems are, basically, optimization problems. The goal is to discover the best among all possible collections of objects according to the given clustering quality function. The most prominent issue of classification text documents is to reproduce the internal structure. Generally, we follow some steps: i) select a measure of similarity; ii) decide on the type of clustering techniques to be used; iii) interpretation of cluster solution.

In text mining, pre-processing of corpus convert unstructured textual data into a structured data. The most common way of doing this is to represent documents by bag-of-words. In this approach, we assume that each word is a dimension in the feature space (Balbi and Di Meglio, 2004). Based on the pre-processing step, we can establish links between statistical units either by using co-occurrence information (within some lexical unit) or by using the semantic relationships between the statistical units. Link recognition based on a process of building up networks of interconnected objects through various relationships in order to discover patterns and trends.

We use one-mode and two-mode social networks (Wasserman and Faust, 1994; Scott, 1997) to analyse the centrality of scholars and research topics, to explore the links between scholars and JADT conferences and research topics and scientists, to visualize the trend of JADT research topics. We classify the papers presented in various editions of JADT, focusing on which techniques are tested and if there have been substantial changes over time. We apply PAM algorithm to search for *k* representative objects or medoids (scholars and research topics) among the objects of the dataset. These objects should represent the structure of the data (Theodoridis S and Koutroumbas K., 2006).

This paper is organised as follows: section 2 describes the link between the Social Network Analysis (SNA) and Textual Mining, introducing the notation and indices used to analyse the micro and macro collaboration structures. Section 3 shows the main results of this approach on the corpus of papers JADT Conferences. Finally, section 4 provides a discussion on previous results and open perspectives.

2. Text Mining vs Social Network Approach

Let a **C** corpus based on a collection of documents, the pre-processing transforms the natural language of **C** from an unstructured representation into a structured representation. However, given the potentially large number of words, phrases, sentences, typographical elements, we obtain a **X** sparse matrix. In **X** matrix, only a small percentage of all possible features for a document collection as a whole appears in any single document, and thus when a document is represented as a binary vector of features, nearly all values of the vector are zero (Stoer and Bulirsch, 2002). Let **X** be a term-document matrix, where the rows correspond to words and columns correspond to documents. The entries may be binary or frequency counts. Similarity between words, usually, is measured by using cosine distance (Iezzi, 2009). Words are linked to each other and together they express concepts that can take on a different meaning untied. If the same words are used in different documents, those documents will most likely have concepts in common and the authors of those documents, perhaps, will have links. The methods suitable to relational data are those of network analysis. Social Network Analysis (SNA) is a set of techniques devoted to analyse relationships between people and/or groups as the most important aspect consisting of nodes and ties (Wasserman and Faust, 1994). Nodes are the individual actors within the networks, and ties are the relationships between actors. Since 1950, SNA is characterised by adopting mathematical techniques especially from the graph theory (Gibbons, 1985; Krackhardt, 1994). A graph, representation of a network, is a pair G=(V, E) consisting of two sets: a set of nodes V={1, 2, …, n} and a set of link E=={e_1, e_2, …, e_k} between pairs of nodes. In the statistical traditional analysis, the data consist of a rectangular array of measurements. The rows of the array are the cases, or subjects, or observations. The columns consist of scores (quantitative or qualitative) on attributes, or variables, or measures. In SNA, we have many ways to represent affiliation networks: 1. Affiliation network matrix; 2. Bipartite graph or Sociomatrix; 3. Hypergraph; 4. Simplicial Complex. Each of these representations contain exactly the same information, and, as a result, any one can be derived from the other.

We suppose, e.g., to analyze four lexical units (1, 2, 3, 4) extracted by five papers (A, B, C, D, E). We could build a case by affiliation matrix to visualize the links between keywords and papers. This is a two-mode network consisting of two sets of units. Several two-mode networks exist: members in institutions, e.g., scientists who participate in several conference on a focus; co-authorship networks - authors, papers, is a (co)author. Figure 1 shows a case by affiliation matrix and a two-mode network: circle points represent the keyword (node-actor) and square points are the papers (node-affiliation).

	AFFILIATIONS				
CASES	**A**	**B**	**C**	**D**	**E**
1	1	1	1	1	0
2	1	1	1	0	1
3	0	1	1	1	0
4	0	0	1	0	1

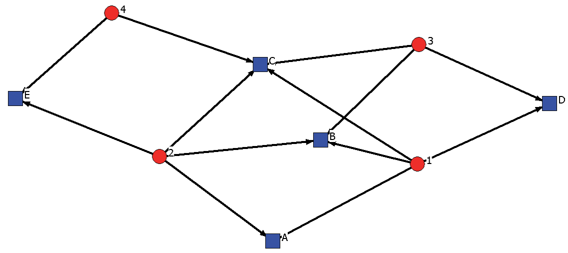

Figure 1: A case by affiliation and graph

A very common and very useful approach to two-mode data is to convert it into two one-mode data sets, and examine relations within each mode separately. We could create a data set of actor-by-actor ties, measuring the strength of the tie between each pair of actors by the number of times that they contributed to the same paper (Fig. 2).

	CASES			
CASES	**1**	**2**	**3**	**4**
1	–	3	3	1
2	3	–	2	2
3	3	2	–	1
4	1	2	1	–

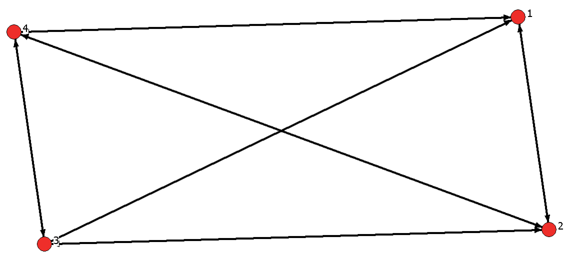

Figure 2: Adjacency matrix (case by case) and graph

We could also create a one-mode data set of paper-by-paper ties, coding the strength of the relation as the number of donors that each pair of initiatives had in common (Fig. 3).

Substantive applications of affiliation networks focus on just one of the modes, such one mode analyses use matrices derived from the affiliation matrix of the graphs defined by such matrices and the affiliation network data is processed to give the ties between pairs of entities in one mode based on the linkages implied by the second mode.

AFFILIATIONS	A	B	C	D	E
A	–	2	2	1	1
B	2	–	3	2	1
C	2	3	–	2	2
D	1	2	2	–	0
E	1	1	2	0	–

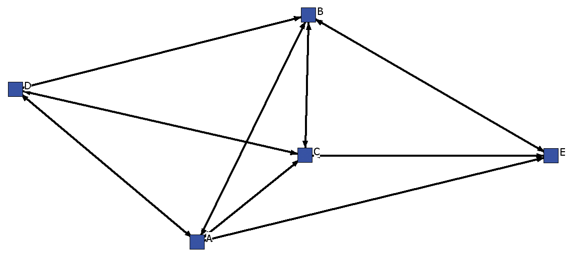

Figure 3: Adjacency matrix (affiliation by affiliation) and graph

Links between lexical units and papers are represented by a graph undirected: two nodes represent two lexical units or two papers and using the same lexical unit relationship creates links. When we study contents of documents published in the closed group, e.g. proceeding of different conferences in JADT editions, we can detect the pattern of cooperation in a macro and micro perspective. In the first viewpoint, graphs inform us about the social structure, where scholars come from and what their research topics and applied statistical methods are, country by country and JADT by JADT. In the second point of view, we analyzed if some scholars and research topics are bridges and, then, if they were deleted, they would disconnect net.

In this approach, we match text mining and SNA methods, in a chain process: i) pre-processing (extraction of relevant lexical unit); ii) relational data matrix definition (affiliation matrix: papers associated to lexical units); iii) attribute data matrix: definition for papers (e.g. authors, position, affiliation,…).

To explore the pattern of word-networks and to describe the collaboration structure between scientific domains, we propose the following sequential steps:

1. Network typologies: graphs;
2. Density, centrality indices and cohesion measures (Wasserman, Faust, 1994);
3. Clustering. (Doreian, *et al*. 2005);
4. Paper-oriented model for longitudinal relational data (Snjiders, 1996, 2001, 2006)

We applied a variety of centrality measure to investigate different contribution of nodes in this networks, which are important to understanding power, relevance in the scientist community and in research topics. These measures give us insight into the various roles and groupings in a network -- who are the connectors, mavens, leaders, bridges, isolates, where the clusters are and who is in them, who is in the core of the network, and who is on the periphery. In particular, we measure degree, closeness, betweenness and eigenvector measures.

Closeness centrality approaches emphasize the distance of a paper to all others in the network by focusing on the distance from each paper to all others. Depending on how one wants to think of what it means to be “close” to others, a number of slightly different measures can be defined. The eigenvector approach is an effort to find the most central papers in terms of the “global” or “overall” structure of the network, and to pay less attention to patterns that are more “local.” The method used to do this (factor analysis) is beyond the scope of the current text. In a general way, what factor analysis does is to identify “dimensions” of the distances among actors. The location of each actor with respect to each dimension is called an “eigenvalue,” and the collection of such values is called the “eigenvector.” Usually, the first dimension captures the “global” aspects of distances among actors; second and further dimensions capture more specific and local sub-structures. The betweenness centrality measure we examined above characterizes papers as having positional advantage, or power, to the extent that they fall on the shortest (geodesic) pathway between other pairs of papers. The idea is that papers that are "between" other papers, and on whom other papers must depend to conduct exchanges, will be able to translate this broker role into power.

In text mining, documents are represented in a vector space information retrieval system (Iezzi, 2009). The document set comprises an **A** (*m* x *n)* term-document matrix, in which each column represents a document, and each entry **A**(*i*, *j*) represents the weighted frequency of term *i* in document *j*. A major benefit of this representation is that the algebraic structure of the vector space can be exploited. To achieve higher efficiency in manipulating the data, it is often necessary to decrease the dimension radically. Especially when the data set is enormous, we assume that the data have a cluster structure, and it is often necessary to cluster the data first to utilize the tremendous amount of information in an efficient way. Once the columns of **A** are grouped into clusters, rather than treating each column equally regardless of its membership in a specific cluster, as is done in the singular value decomposition (SVD), the dimension reduction methods we discuss attempt to preserve this information. In SNA, «Blockmodels are used to collapse redundant elements in a system in order to clarify the patterns of relationships among the elements» (Borgatti and Everett, 1992). Blockmodeling is based on the idea that permuting the rows and columns of a matrix we can reveal a pattern (Breiger et al., 1975). We apply several clustering algorithms (*k*-means, DIANA, PAM, CLARA, FANNY, SOM, SOTA and Model Based-clustering) to select clusters of scholars and research topics. In order to validate internal consistency of groups, we apply Connectivity, Silhouette Width and Dunn index and to measure stability we apply Average proportion of non-overlap, average distance and average distance between means (Darra and Datta, 2003; Handl et al, 2005; Brock et al., 2008).

3. JADT Conferences

International Conference on statistical analysis of textual data conferences (JADT) is a biennial meeting among scholars and researchers working in the field of textual data analysis. The main topics range from lexicography to the analysis of political discourse, from information retrieval to marketing research, from computational linguistics to sociolinguistics, from text mining to content analysis. Since 1990 there have been nine editions: Barcellona (1990), Montepellier (1993), Rome (1995), Nice (1998), Lousanne (2000), St. Malo (2002), Louvain (2004), Besancon (2006), Lion (2008) and the last to be held in Rome (2010). France has hosted five editions (1993, 1998, 2002, 2006, 2008), Italy two (1995 and the last in 2010), Spain (1990), Switzerland (2000), Belgium (2004) how hosted one Conference. In twenty years, 1642 researchers participated in writing a contribution to the JADT (Fig. 4). They come from 34 different countries and belong to the five continents, although the largest share of researchers come from Europe, particularly from France (43.74%) and Italy (22.23%). A significant number of scientists came from Canada (9.31%), Belgium (6.83%), Spain (4.14%) and Suisse (2.69 %). A small group of researchers come from the UK (1.45%), Germany (1.14%), Japan (1.14%), the Usa (0.93%), Algerie (0.83%), Iran (0.62%), Russia (0.62%), Tunisia (0.62%), Argentina (0.52%), Mexico (0.52%), The Netherlands (0.41%), India (0.31%), China (0.21%), Japan (0.21%), Portugal (0.21%), Australia (0.10%), Brasil (0.10%), Checz (0.10%), Cuba (0.10%), Egypt (0.10%), Finland (0.10%), Irlande (0.10%), Marocco (0.10%), Romania (0.10%), Saudi Arabi (0.10%), Siria (0.10%), United Arab Emirates (0.10%) and Zimbabwe (0.10%).

In different editions, the papers presented at JADT have increased over time, passing from 17 papers published in Barcelona (1990), to 136 papers in Rome (Fig 5). The median of papers was 80.

Figure 1 shows a steady growth in the number of papers presented, three editions are particularly rich in contributions: Rome (1995 and 2010), Louvain (2004) and Lyon (2008), in which the work submitted were near or exceeded 100. On average, a paper has a size of 10 pages.

The ratio between number of authors and papers presented at JADT is a rough measure of collaboration on each work publication. The average number of teamwork is 1.92 (SD 0.30) researchers per paper published.

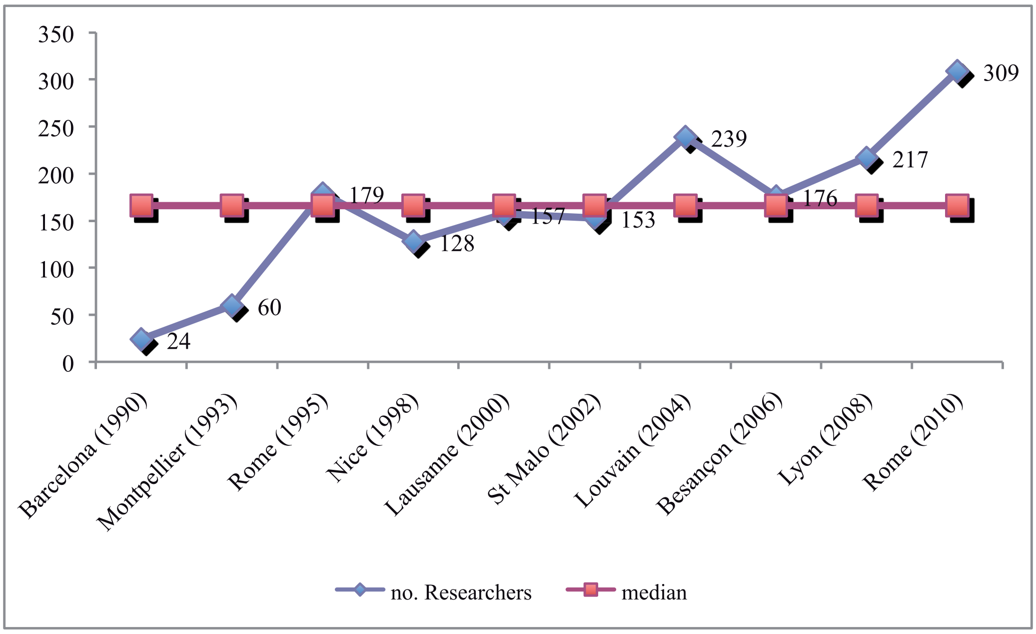

Figure 4: No. Researchers at JADT Conferences (1990- 2010)
Source : Elaborations on proceedings of JADT Conferences (1990-2010)

The highest value occurs in Rome (2.27 researcher per paper), in St. Malò (2.2 researcher per paper) and the lowest in Montepellier (1.33 2 researcher per paper) .

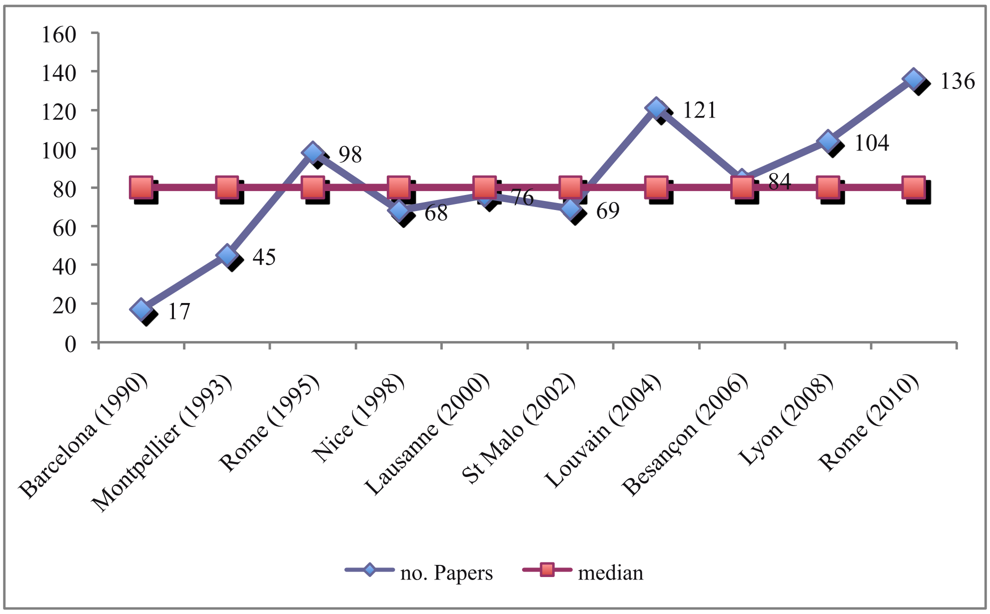

Figure 5: No. Papers presented at JADT Conferences (1990- 2010)
Source : Elaborations on proceedings of JADT Conferences (1990-2010)

The number of scientists present in the various editions JADT (Fig. 6) is strongly correlated with the number of papers submitted ($\rho = 0.98$).

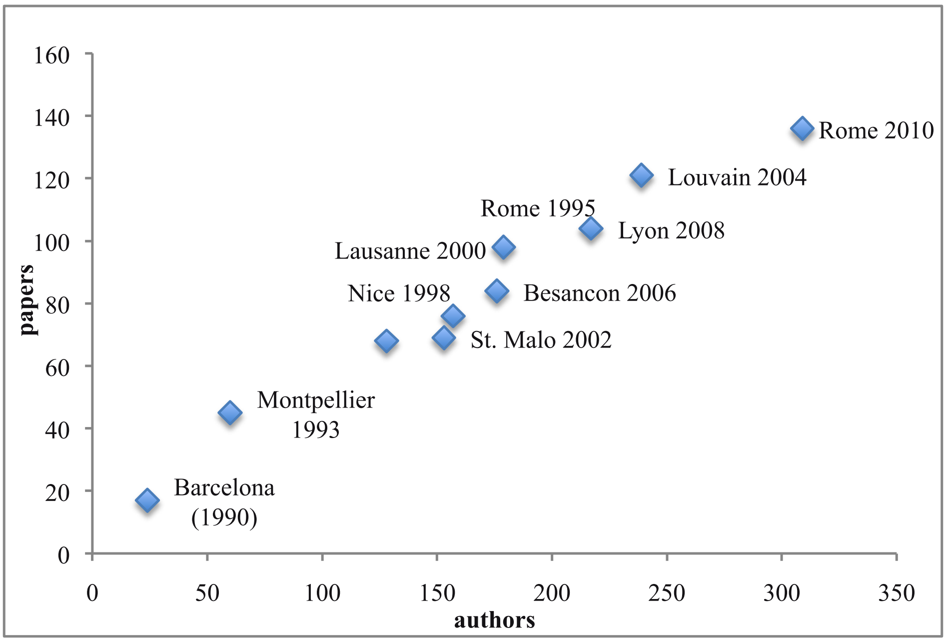

Figure 6: Papers published by authors
Source : Elaborations on proceedings of JADT Conferences (1990-2010)

The corpus is composed by 818 papers. The papers were written in six different languages: French (59.6%), English (25.3%), Italian (11.7%), Spanish (2.2%), Catalan (1.1%) and German (0.1%). In all editions, the prevailing language is French, although there were many papers written in English and Italian (Fig. 7). In the last edition, 47.5% of the papers were written in English.

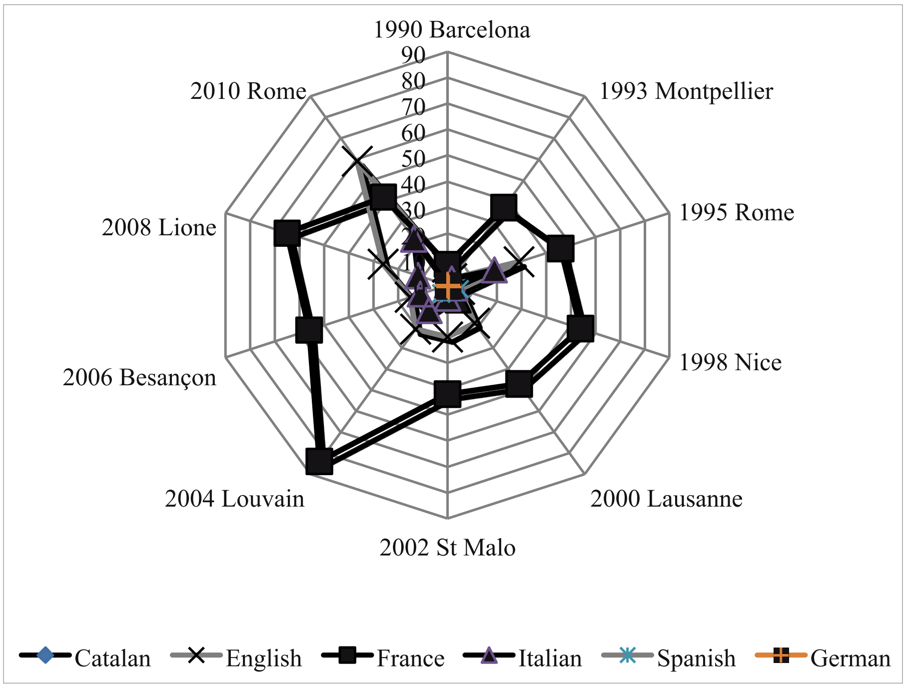

Figure 7: Language of papers (%)

All abstracts, titles and keywords (if included) were translated in English to classify the contents of the papers. The corpus of abstracts is composed by word token no. 147,206 and word type no.39,745. The titles present about 10 words to each paper. Using measures of centrality, we are able to identify scholars who occupy more "influential" positions in the JADT conferences. Table 1 shows the centrality of the most important scientists of JADT. We can see those nodes are similar in cooperating with many members of JADT.

Degree centrality is simply the number of direct relationships that an entity represents. High degree of centrality for an actor generally corresponds to lots of papers published at JADT conferences. Generally, he has an active role in the organization of this conference, e.g., he could be a member of the scientific committee. Table 1 shows the nodes that are connectors or hubs in the network. Closeness centrality measures how quickly an entity can access more entities in a network. An entity with a high closeness centrality generally: has written many papers with other group member. Betweenness centrality identifies an entity's position within a network in terms of its ability to make connections to other pairs or groups in a network. Eigenvalue measures how close an entity is to other highly close entities within a network. In other words, eigenvalue identifies the most central entities in terms of the global or overall makeup of the network. Labbé, e.g., has participated actively in almost all the conferences and has established a scientific collaboration with many other authors. These characteristics are in common in all the authors listed in Tab. 1.

Node	*degree*	*betweenees*	*Closeness*	*Eigenvector*	*k-kore*
Labbé	9	17.144.277	4.140.000	0.136	1.455.000
Lelu	8	13.261.507	4.368.000	0.116	1.258.000
Areni	8	13.875.040	4.196.000	0.133	1.398.000
Balbi	8	13.875.040	4.196.000	0.133	1.398.000
Sensales	8	13.875.040	4.196.000	0.133	1.398.000
Beaudouin	8	14.084.982	4.292.000	0.120	1.295.000
Marchand	7	7.165.831	4.890.000	0.075	852.000
Bolasco	7	7.433.612	4.834.000	0.085	897.000
Meunier	7	9.130.461	4.394.000	0.122	1.238.000
Viprey	7	9.130.461	4.394.000	0.122	1.238.000
Brunet	7	10.168.566	4.508.000	0.105	1.119.000
Duchastel	7	11.673.400	4.552.000	0.093	1.036.000
Reinert	6	4.591.998	5.096.000	0.059	692.000
Bécue	6	5.697.305	4.812.000	0.087	939.000
Savoy	6	6.220.948	4.532.000	0.113	1.124.000
Turchi	6	6.220.948	4.532.000	0.113	1.124.000
Martinez	6	6.357.595	4.652.000	0.100	1.016.000
Lebart	6	6.954.983	4.820.000	0.081	877.000
Biskri	6	7.053.876	4.534.000	0.109	1.098.000
Luong	6	8.055.459	4.582.000	0.097	1.036.000
Moscarola	5	3.065.304	5.178.000	0.055	634.000
Salem	5	4.258.002	5.010.000	0.068	742.000

Table 1: Centrality of major nodes

French and Italians have established an extensive network of collaborations (Fig. 8). In contrast, the isolated nodes are scholars who attended less than three conferences. Not with standing this, there were many isolated countries at the Rome conference (1995) and at the Lyon conference (2008). Australia, Brasil, Checz, Cuba, Egypt, Finland, Irlande, Marocco, Romania, Saudi Arabi, Siria, United Arab Emirates and Zimbabwe are countries not integrated into the network of work JADT. These scientists are probably also attracted by the conference venue.

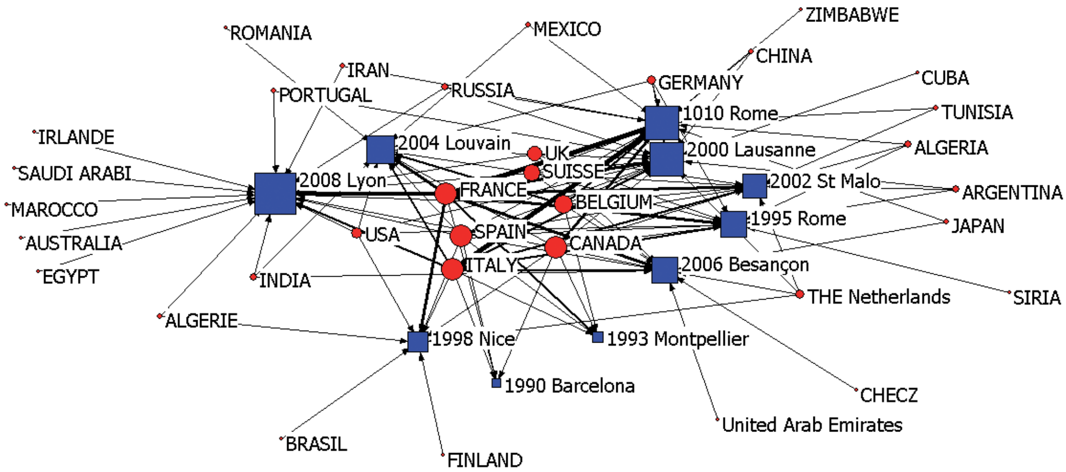

Figure 8: The cooperation graph of JADT (degree centrality)

The map (Fig. 9) shows that, in the first edition, scientists presented methods of classification and applications in education. At the second and third editions, the papers dealt with open-ended questions and issues of hyperlinks. They range from issues concerning the analysis of repeated segments to technical problems of tagging and multilingualism. What seems quite prominent, even from an analysis above all abstracts, is that, in early editions, there are a lot of statistical contributions, while, in later editions, although the focus on these issues increases, in terms of scientists who participated, but multidisciplinary was privileged. In recent editions, there are few papers that present new statistical methods. There are many papers with applications in different fields of research.

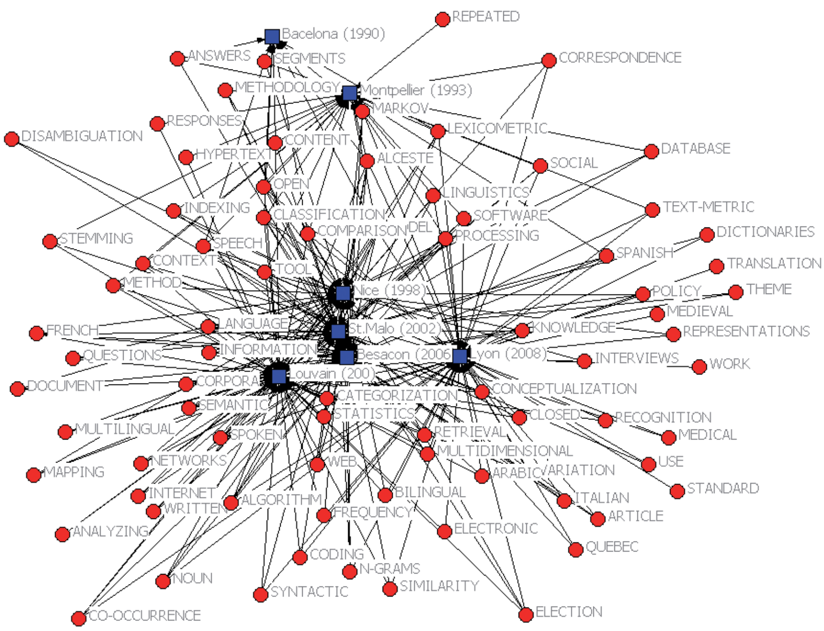

Figure 9: The cooperation graph of papers

Fig. 8 shows a high closeness centrality based on origin of scientists. France, Italy, Spain, Canada and Belgium are more connected.

We have chosen a "bag-of-the-words" representation, in which each key-topic is represented as a vector consisting of all the unique words and bigrams (pairs of consecutive words) appearing in the topics and the number of times each appears. We ignored all words that do not appear at least five times.

By running the analyses, we applied UCINET 6.0 and R software.

4. Conclusions and Future steps

In this paper, we presented some possible approaches to analyse the network of scientists and affiliation research topics. It is interesting to note that JADT strongly depends on few nodes (Statistician), but these do not affect most of the topics covered. The majority of papers come from individuals not related to the central nodes. In 20 years, JADT has created a large network, composed of 1642 scientists, although only about thirty are very active. Many authors have participated in a single edition, without creating links with other authors. These are mainly scholars who have submitted applications in the field of psychology, sociology and linguistics. Over time, JADT has increased the number of papers published, but has changed the core works. It expanded the range of applications, while the space reserved for new statistical methods has been reduced. In the first editions, there were many papers of multivariate statistical techniques: classifications, multidimensional scaling and factorial methods. In the later editions, JADT devoted applications to text mining.

It has become the privileged place where scholars from different fields gather together to share methods and applications. JADT has helped to identify a common language, even if the works are presented in many different languages. This issue is a real strength, since it is not uncommon that the same methods are used in diverse areas, but in a different language. The exchange of views facilitates the ability to create connections and enrich scientific knowledge. Indeed with the advent of the Internet, online resources have been increasingly available. Many users choose popular search engines to perform an online search to satisfy their information needs. Many links could be among scientists of different disciplines and it is possible to find interesting new hypotheses of study. Literature based discovery (LBD) aims to find unpublished relationships between domain entities (Gordon *et al*., 2002) in a virtual place, JADT meets scientists in a physical place. It is a closed system, in which we can discover many indirect relationships between scientists. In the future, we will apply SNA of textual methods in an open system through search engines.

References

Balbi S. and Di Meglio E. (2006). A Text Mining Strategy based on Local Contexts of Words, In *Proceedings JADT - 2004 7èmes Journées internationales d'Analyse statistique des Données Textuelles*. Louvain-La-Neuve, Belgium, Vol **1**, pp 79-87.

Batagelj V. (1997*).* Note on blockmodeling. In *Social Networks*, **19,** pp. 143-155.

Bolasco S. (2005). Statistica testuale e text mining: alcuni paradigmi applicativi, in *Quaderni di Statistica*, Liguori Ed., **7**, pp. 17-53.

Borgatti, S. P., & Everett, M. G. 1992. Regular blockmodels of multiway, multimode matrices. In *Social Networks*, **14**: 91-120.

Breiger, R. L., Boormam,, S. A., and Arabie, P. An algorithm for clustering relational data, with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology*, (1975). **12**, 328-383.

Brock G, Pihur V, Datta S, Datta S: clValid, an R package for cluster validation. *Journal of Statistical Software* 2008, 25:1-22.

Datta, S. and Datta, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics* 19(**4**): 459-466.

Doreian, P., Batagelj, V., & Ferligoj, A. (2005). *Generalized Blockmodeling*, Cambridge University Press.

Feldman R., Sanger J. (2007). *Text mining handbook*. Advanced Approaches in Analizing Unstructured Data, Cambridge Press.

Gibbons A. (1985). *Algorithmic Graph Theory*, Cambridge University Press.

Gordon A.D. (1999). *Classification*. Chapman & Hall CRC.

Gordon M, Lindsay R.K, Fan W. (2002). Literature-based discovery on the World Wide Web, in *ACM Transactions on Internet Technology*, Vol 2/4, pp. 261-275.

Handl, J., Knowles, K., and Kell, D. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics* 21(**15**): 3201-3212.

Hanneman R. and Riddle M. (2005). Introduction to social network methods. (published in digital form at Available at http://www.springerlink.com/index/ http://faculty.ucr.edu/~hanneman/)

Iezzi D.F. (2009). Intimate femicide in Italy: a model to classify how killings happened. In Palumbo F., Lauro C.N. and Greenacre M.J., editors, *Data Analysis and Classification*, Berlin: Springer-Verlag, pp. 85-92.

Lee, S. Bozeman, B. (2005). The impact of research collaboration on scientific productivity, In *Social Studies of Science*, 35 (5), 703-723.

Krackhardt. D. (1994) Graph Theoretical Dimensions of Informal Organizations, In *Computational Organization Theory*, pp 89-111,

Snijders, T.A.B. (1996) Stochastic actor-oriented dynamic network analysis. In *Journal of Mathematical Sociology* 21, 149–172.

Snijders, T.A.B. (2001) The statistical evaluation of social network dynamics. In: Sobel, M., Becker, M. (Eds.), *Sociological Methodology*, Basil Blackwell, Boston and London, pp. 361–395.

Snijders, T.A.B., Steglich, C., Schweinberger, M., and Huisman, M. (2006) *Manual for SIENA version 2.4*. Groningen: ICS, University of Groningen.

Stoer J., Bulirsch, R. (2002). *Introduction to Numerical Analysis* (3rd ed.). Berlin, New York: Springer-Verlag. ISBN 978-0-387-95452-3.

Theodoridis S, Koutroumbas K. (2006). *Pattern Recognition* 3rd ed..

Wasserman and Faust K. (1994). *Social Network Analysis: Methods and Applications*, Cambridge University Press.

Van Rijnsoever F. J., Hesselsand L.K., Rens L ,Vandeberg.J. (2008). A resource-based view on the interactions of university researchers. In *Research Policy*, Vol. 37, Issue 8, September 2008, pp. 1255-1266.