Basic principles of bibliometrics.
Application to research development

Hervé Rostaing
Aix-Marseille University
Scientific Center of Saint Jérôme
F-13397 Marseille Cedex 20
FRANCE
Tel: +33 4 91 28 87 46
Fax: +33 4 91 28 87 12
e-mail: herve.rostaing@univ-cezanne.fr
web: http://crrm.u-3mrs.fr

Introduction: definitions, history and fundamentals of bibliometrics

Bibliometrics is a set of techniques for analyzing large amount of bibliographic data using mathematical, statistical and computational methods. These quantitative analyses are used to discover relationships, trends and models that represent the evolution and the construction of science and technological rises.

The first bibliometric studies started at the beginning of the twenty-century. These first studies were especially focus on the mathematical and statistical analysis of data distribution. Some of these studies marked the scientific community at such a point that the researchers who undertook these studies gave their name with certain bibliometric "laws". These bibliometrics “laws” try to formulate statistical data distributions mathematically. Some of these studies influenced the scientific community so much that the researchers at the origin of these studies gave their name with certain bibliometric "laws". These bibliometrics laws try to formulate statistical data distributions mathematically: Zipf law (Zipf, 1949), Lotka law (Lotka, 1926), Bradford law (Bradford, 1948)…

Thereafter, in the Sixties, De Solla Price was one of the key figures of a new field of application of these bibliometric techniques. It was based on these quantitative studies to consolidate its sociological theories for the study of science (Price, 1963). From this school of thought, were born the data bases of ISI[1] (SCI, SSCI, A&HCI) like essential tools for the realization of its experiments. A new terminology appeared to define this new field of application of these techniques: *Scientometrics* or called more eulogistically *Science of science*.

During this period, the first science mapping experiments were born (Small, 1973). The creation of these maps finally made it possible to represent the relations maintained between the bibliographical data. These maps try to visually recreate the cohesion of the scientific works and the structure of the relations established between this works.

[1] Institute for Scientific Information http://www.isinet.com created by Garfield (Garfield, 1979)

In the Eighties, the democratization of computers, the use of telecommunications to connect computers and thus the access to the large scientific and technical databases allowed a broader diffusion of the application of these bibliometric techniques. Many actors could try out these techniques on their own data and some new fields of application emerged.

Competitive intelligence and bibliometrics

Since few years, bibliometric techniques have found new applications directed closely towards company needs. Bibliometrics is especially suitable for supporting a Competitive Technical Intelligence process for companies.

The Competitive Intelligence tracks all activities of direct or indirect competitors. The Competitive Technical Intelligence is focused on information useful for research and development activities of competitors. The constant growth of textual data sources, the constant growth of the complexity for understand the fragmentation of these textual data prompt the companies to integrate automatic analysis of textual data. Bibliometrics techniques can play an important role for exploiting this amount of textual data as well as possible.

Competitive technological Intelligence (Technology Watch) is focus on the tracking of scientific, technical and technological information. The collected textual data on scientific, technical and technological activities of competitors have to be transformed into information and intelligence to help policymakers. In general terms, the intelligence cycle can be defined as the process by which raw data is acquired, gathered, evaluated, analyzed and transmitted as finished intelligence to decision-makers. There are five steps that constitute this cycle: planning and direction, targeting and collection, processing and analysis, production and dissemination, and decision and action

In this cycle, if we consider that the step of targeting the sources and collecting the textual data is on control by the contribution of current communication and information technologies (NICTs) and the planning of informal information collecting (what is not so obvious) then the step of processing and analyzing the collected data becomes critical to guarantee the success of the intelligence cycle.

Very often, when the activity of competitive technical intelligence is upstream from the launching of a project of innovative development, a very great number of knowledge still absent or ignored in the company must be evaluated and assimilated. In this type of competitive technical intelligence, the volume of collected data to treat is very huge and can be very confusing. Then it becomes difficult to ask experts to carry out their works of evaluation, filtering, interpretation and analysis without providing a sustainable help. To make completely efficient their action and to make sure of a continuous motivation in this action, it is vital that they feel supported and accompanied.

The bibliometric techniques will be very useful for this step of knowledge structuring and supported process for experts. Many examples carried out by the CRRM lab in French companies proved the benefit of bibliometric treatments for the competitive technical intelligence: Bisson (2003) in Automatech an SME in circuit board industry; Liège (2003) in Danone Vitapole the R&D center of the international food company; Da Silva (2002) in Snecma Motors an aeronautical and space motorization company; Catapano (2001) in CLL.Pharma an SME in generic drugs development; Lauri (1998) in Gemplus a leader smartcard company, Dumas (1994) in CETIM a R&D center of mechanical industry; Nivol (1993) in L'Oréal a leader cosmetic company.

Bibliometric treatments and textual data processing protocol

The bibliometric treatments in competitive technical intelligence have the aim of offering to the experts a quick "reading grid" of great volumes of textual data. This reading grid enables them to apprehend a greater number of texts than by simple reading. The global vision offered by the bibliometric results helps to structure the new field of knowledge contained in these texts using graphical visualization (curved, histograms, networks of relations, charts…).

To achieve this goal, the bibliometric treatment follows five main steps:

- Textual data collection (the corpus studied)
- Corpus division in statistical units
- Extraction and/or determination of the elements describing these statistical units
- Graphical visualization of the statistical result

Textual data collection : the corpus

Bibliometrics analyses use mainly three main sources of documents: patents references, references of scientific articles with citations[2] and references of scientific articles with no citations.

Corpus division in statistical units

For any statistical analysis of a corpus of textual data, it is necessary to choose the elementary unit that will be the object of the analysis: the statistical unit.

For the analysis of full texts (literary works or political speeches…) the statistical unit is often the sentence, the paragraph or the chapter, the section or a window of a definite number of words.

For bibliometric analysis the natural statistic unit is the bibliographic reference. Thus, in the case of scientific corpus of scientific articles the elementary object analyzed is the description of a scientific work. While in the case of a corpus of references patents, the elementary object is the description of a technological innovation protected by a patent.

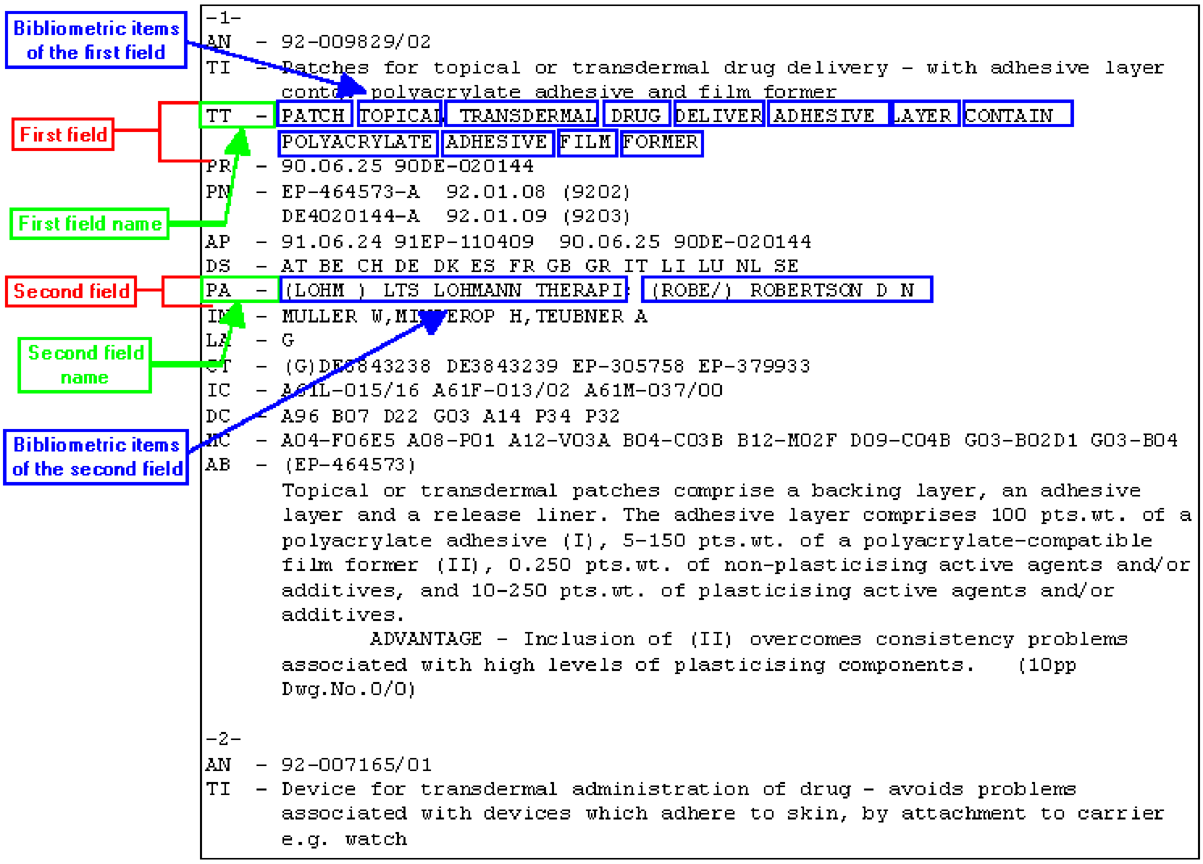
[2] Databases produced by ISI contain article references including the bibliography mentioned in each article.

Extraction and/or determination of the elements describing these statistical units: the characteristic elements

These statistical units are the subject of statistical processing which aim is either to make comparisons between units or make regroupings of units per similarity or by resemblance. Thus these statistical objects must be characterized for classifying, comparing or grouping treatments.

Generally these characteristic elements result directly from the bibliographic references themselves by an automatic process of information extraction. This process of information extraction is facilitated because the bibliographical references from the databases are available with a very structured format. Each piece of data is preceded by a field label giving the nature of the piece of data (Fig 1). A field can contain more than one value of data of the same nature. A multi-variant field separates each value by a separation character (comma, semi-colon, space…).

Thus, each statistical unit can easily characterized by the writers of the work (authors or inventors), the organizations employing these writers (laboratories or companies), the countries where these authors come from (or countries covered by the invention), the year of the publication, the name of the journal publishing the work and the concepts approached in the work (words, keywords, codes).


Bibliometric items of the first field
First field
First field name
Second field
Second field name
Bibliometric items of the second field

```
-1-
AN  - 92-009829/02
TI  - Patches for topical or transdermal drug delivery - with adhesive layer
      contg. polyacrylate adhesive and film former
TT  - PATCH TOPICAL TRANSDERMAL DRUG DELIVER ADHESIVE LAYER CONTAIN
      POLYACRYLATE ADHESIVE FILM FORMER
PR  - 90.06.25 90DE-020144
PN  - EP-464573-A  92.01.08 (9202)
      DE4020144-A  92.01.09 (9203)
AP  - 91.06.24 91EP-110409  90.06.25 90DE-020144
DS  - AT BE CH DE DK ES FR GB GR IT LI LU NL SE
PA  - (LOHM ) LTS LOHMANN THERAPI (ROBE/) ROBERTSON D N
IN  - MULLER W,MINTEROP H,TEUBNER A
LA  - G
CT  - (G)DE3843238 DE3843239 EP-305758 EP-379933
IC  - A61L-015/16 A61F-013/02 A61M-037/00
DC  - A96 B07 D22 G03 A14 P34 P32
MC  - A04-F06E5 A08-P01 A12-V03A B04-C03B B12-M02F D09-C04B G03-B02D1 G03-B04
AB  - (EP-464573)
      Topical or transdermal patches comprise a backing layer, an adhesive
      layer and a release liner. The adhesive layer comprises 100 pts.wt. of a
      polyacrylate adhesive (I), 5-150 pts.wt. of a polyacrylate-compatible
      film former (II), 0.250 pts.wt. of non-plasticising active agents and/or
      additives, and 10-250 pts.wt. of plasticising active agents and/or
      additives.
            ADVANTAGE - Inclusion of (II) overcomes consistency problems
      associated with high levels of plasticising components.   (10pp
      Dwg.No.0/0)

-2-
AN  - 92-007165/01
TI  - Device for transdermal administration of drug - avoids problems
      associated with devices which adhere to skin, by attachment to carrier
      e.g. watch
```

Figure 1: Tagged format of a bibliographic reference. This sample of patent references comes from Derwent WPI database.

The characteristic elements obtained by an automatic process of information extraction must very often undergo a treatment of cleanup or grouping of concepts. The cleanup treatment can be defined as the manual or automatic standardization of the characteristic elements by deletion of useless elements or misspelled elements checking. The grouping exercise includes the identification of synonyms and alternate elements for describing a similar concept. These tasks are required in order to produce statistically relevant results otherwise the bibliometric results could be misleading or not truly represent the relative importance of the corpus characteristics.

Graphical visualization of the statistical results

The basic statistical measurement in bibliometric studies is the calculation of the frequency of each characteristic element. The element frequency is the number of references in which this element is present.

The generation of frequency lists provides counts of the various characteristic elements within a bibliographic field and allows a comparison of all these elements. This result is the most common task conducted in bibliometric studies.

These frequency lists can graphically be represented either in histogram form or pie-chart form or in curve form (Fig 2, 3 and 4). The exemples presented below come from a bibliometric study carried out on the Algerian scientific articles published during a 10 years period (1990-1999). This study was exposed at the ISSI'2001 congress (Rostaing et al, 2001).

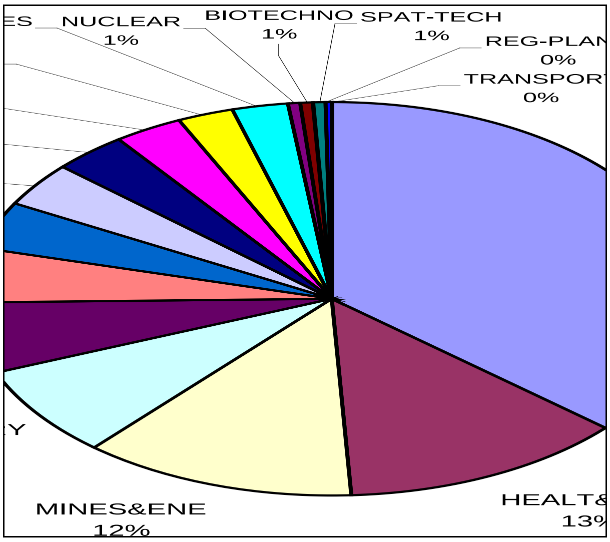


Figure 2: Share of publications for Algerian research fields

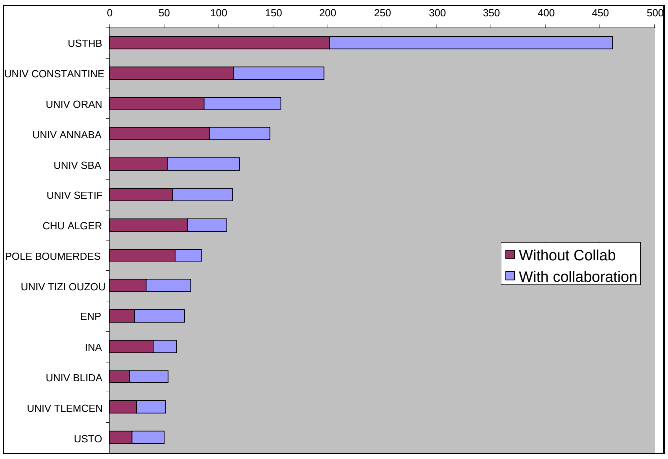


Figure 3: Main Algerian institutes

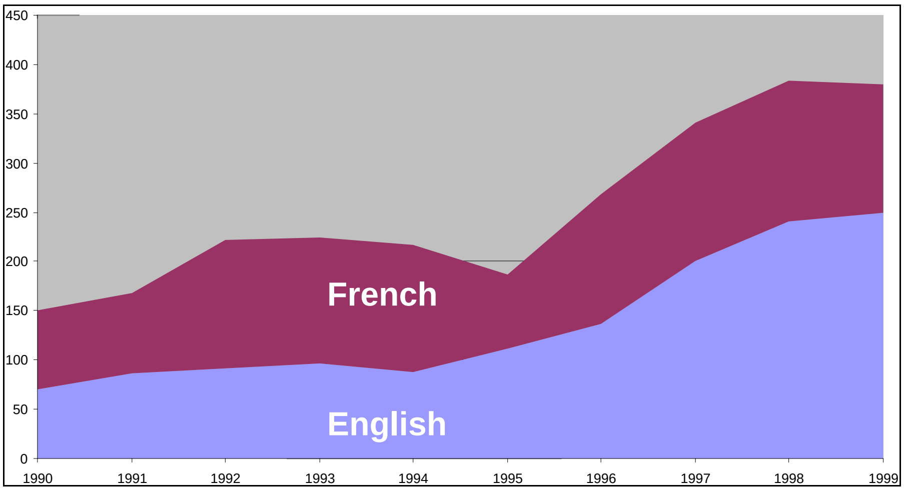


Figure 3: Evolution of Algerian scientific production and the languages used for publishing

These bibliometric results can be issued from all the characteristic elements of the corpus (authors, organizations, dates, journals, countries, classification codes, keywords…). Nevertheless, for most characteristics elements it is strongly recommended that a cleanup and grouping process be applied to them first to help increase the relevance of these frequency lists.

Such uni-dimensional statistical analyses provide many information but do not answer all expectations because the characteristic elements of the corpus are studied separately and because no information on the relations between these elements is offered.

Elements relations analysis

This is the reason why many bibliometrics techniques are based on the study of co-occurrence analysis. The computation of co-occurences of characteristic elements is very often expressed by the generation of co-occurrence matrices. The inventory of the relations between the studied elements is presented in tabular form.

These matrices are built up either to measure the associations between the elements belonging to the same bibliographical field (intra-field relation) or to measure the associations between elements belonging to two or several fields (inter-fields relation). The inter-field matrices are well known under the name of contingency matrix.

These matrices distribute in each row and each column a characteristic element of the corpus coming from one or more fields and the intersection between a row-element and a column-element measure the association between these two elements. The merest measurement of a bibliometric association is the co-occurence frequency of these two elements. This co-occurence frequency corresponds to the number of references where these two elements are simultaneously present. More sophisticate measurements can be computed for reducing the weight of very frequent elements or increasing the weight of infrequent elements. The table 1 shows an example of a co-occurrence matrix using co-occurrence frequency measurement for studying the relations between the scientific fields of Algerian articles and the publishing date. The value at the intersection of a row and a column represent the number of articles published in a scientific field for a year.

	1990	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
FONDA-SCI	60	69	101	90	81	77	106	108	150	164
HEALT&MED	35	31	33	54	47	19	24	42	37	34
MINES&ENE	24	18	36	14	31	26	49	55	40	46
INDUSTRY	2	8	13	10	18	15	27	35	33	49
TECH-INFO	2	5	7	7	2	8	18	30	45	37
INDU-TECH	6	2	5	4	6	9	22	20	26	25
AGRI&FEED	5	16	10	13	8	9	4	5	21	16
ENVIRONME	3	6	10	8	9	12	12	13	12	13
ARID-REGI	9	4	8	3	8	7	13	14	20	10
ENG&TECHN	3	3	2	2	3	4	7	18	18	19
RENEW-EN	3	1	2	17	4	3	9	7	11	14
HYDRO-RES	1	3	3	6	8	2	5	17	15	8
NUCLEAR		1		3		1	3	1	3	7
BIOTECHNO			1	2	1	3	1		4	4
SPAT-TECH		2	1	3		1		3	2	2
REG-PLANN							2		2	
TRANSPORT					1		1			1

Table 1: Co-occurrence matrix between the publication dates and scientific fields of Algerian articles

A graphical representation of such matrices quickly becomes essential for analyzing it. When the size of the matrix is not too large (not too many lines and columns) then the representation in chart form is possible (Fig 4).

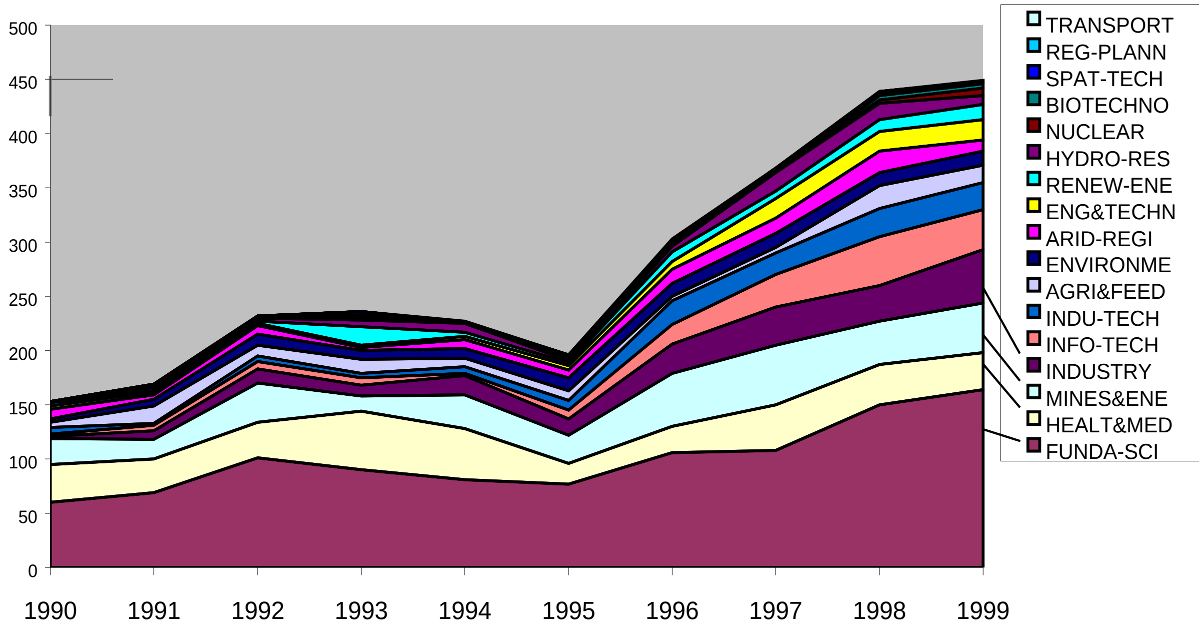


Figure 4: Chart visualization of the co-occurrence matrix of tab 1

Mapping drawing

But very often the matrices generated for bibliometric studies are made up of far too many elements that this style traditional chart could not be easily understandable (see example in Fig 5). Bibliometric techniques use statistical multidimensional methods such as clustering methods (hierarchical classifications, K-means clustering…) or factorial analysis (multidimensional scaling, correspondence analysis…) for contributing to the analysis of complex matrices. These well known statistical methods seek to reduce the complexity of associations of the elements included in the matrix by grouping them by similarity.

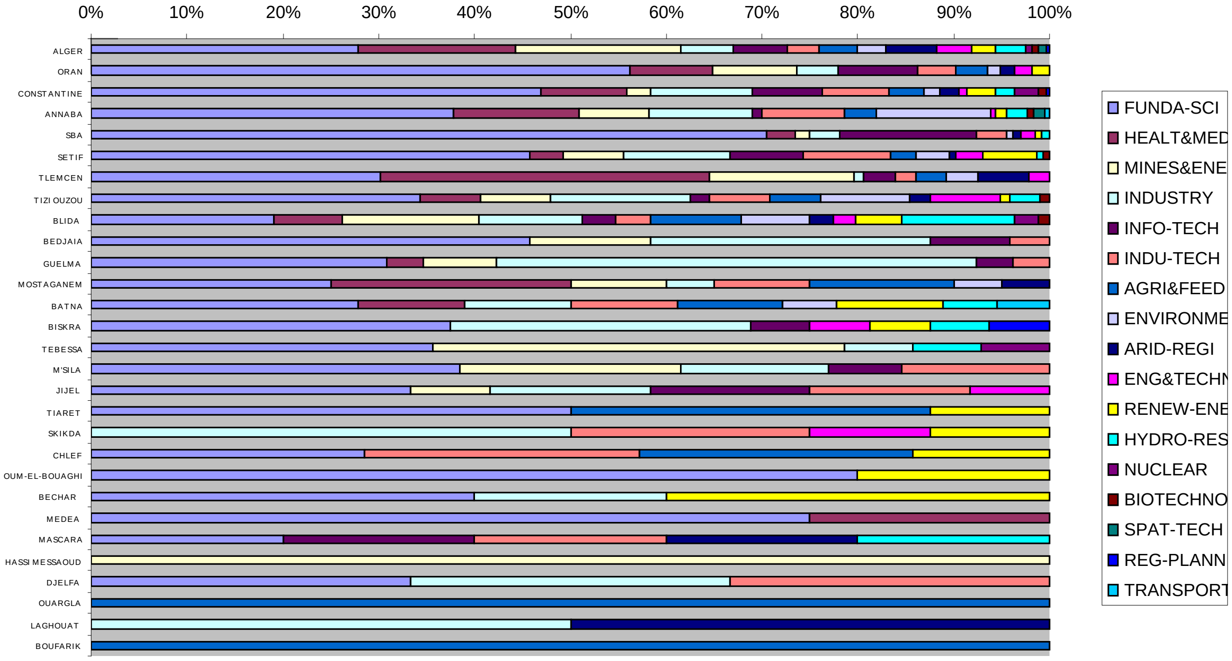


Figure 5: Scientific and technological specialization of Algerian towns

The clustering methods seek to create the most homogeneous groups of elements. Among the hundreds or thousands of possible combinations of element distribution in groups, these methods find a solution optimizing at the same time the similarity of the elements belonging to the same group and the dissimilarity between groups.

The methods of factorial analysis also have the aim of complexity reduction of the relations described in a matrix. These methods privilege a visual representation of associations between these elements by mapping drawing. The elements are represented in the shape of a group of dots that is projected on a plan or in a 3D space. The dots are laid out on a map of manner so that the elements most strongly associated are closest on the map. The user can then set up groups of similar elements by a visual analysis of the map.

Figure 6 show the map obtained with a correspondence analysis applied to the same matrix which was represented by traditional chart in Fig 5. Nevertheless, such a map can easily be misinterpreted by the projection deformation of the dots on space reduced to two dimensions whereas they were calculated in a multidimensional space. This projection inevitably distorts the real distances between the dots and certain dots can be close on the two dimensions map whereas they are not in the multidimensional space. To avoid bad visual interpretations, it is preferable to supplement this map by a clustering method. For achieving this goal, the coordinates of the dots in multidimensional space (calculated by the correspondence analysis) are recorded in a new matrix. A clustering method applied to this coordinates matrix allows to identify groups of the closest dots in multidimensional space. Figure 7 shows the result of this regrouping. According to the hierarchy of the successive regroupings, 9 groups of dots can emerge. The projection of these 9 groups on the initial map (Fig 8) contributes favorably to visual interpretation. If the two-dimension map does not provide a fitting representation to the real positions of the dots, it is possible to prefer a three-dimension representation (Fig 9 and Fig 10).

Some works showed the interest of the use of such bibliometric mappings for helping the comprehension of technical fields (Escorsa et al., 2000) or scientific fields (Kostoff, 1997) in a context of competitive intelligence.

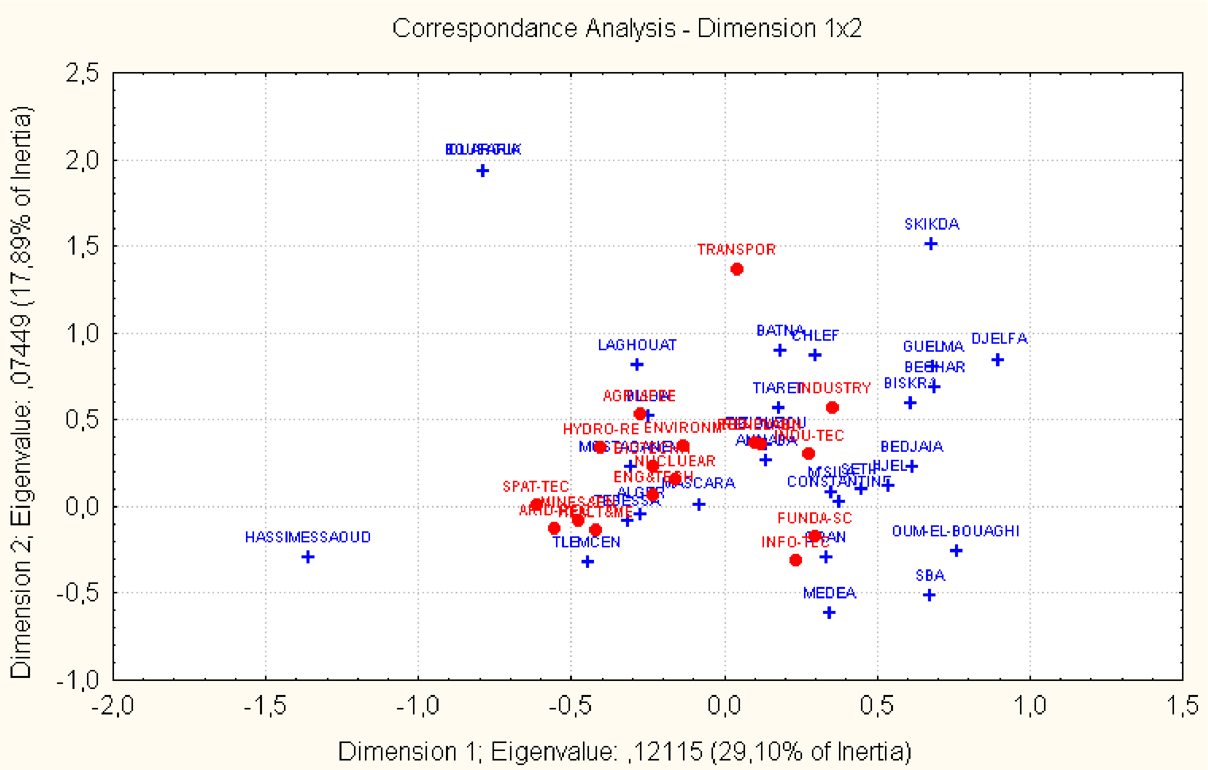


Figure 6: Map obtained with correspondence analysis
of the matrix that was represented in Fig 5

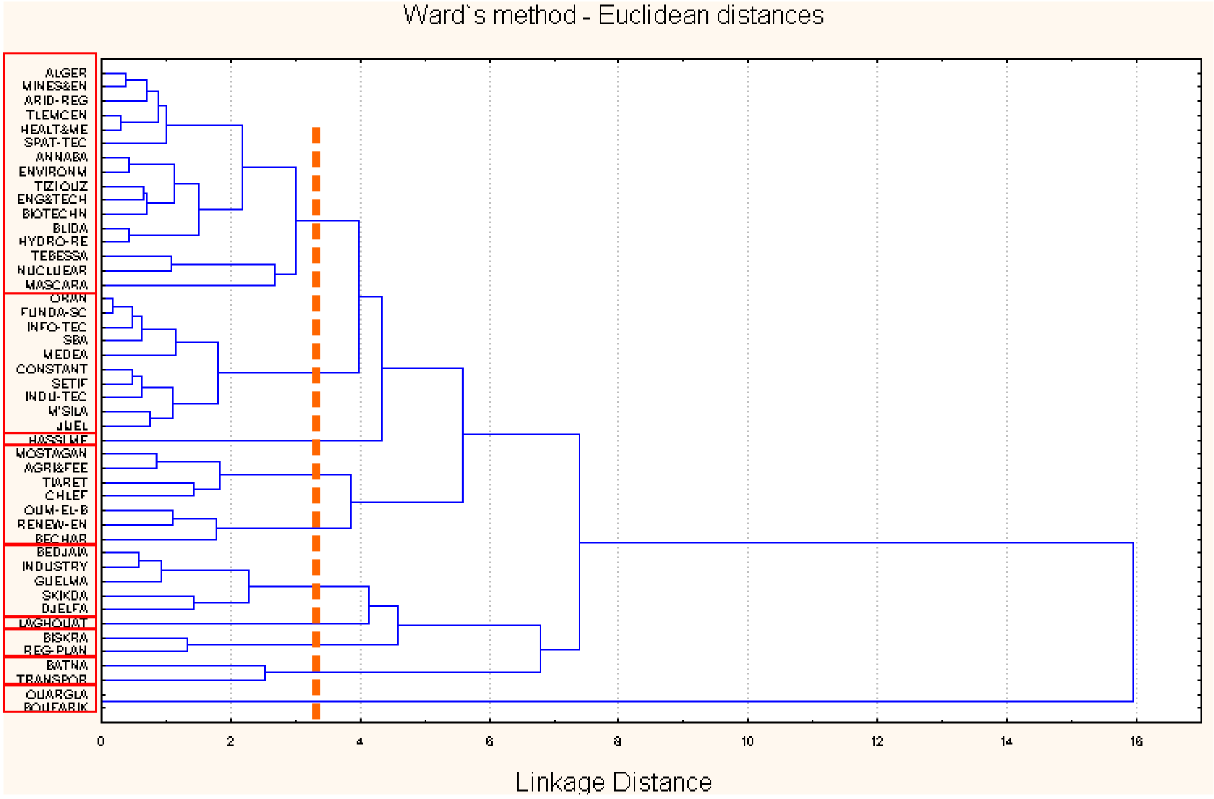


Figure 7: Element regrouping with a clustering method applied on the coordinates of the elements in the multidimensional space of the correspondence analysis

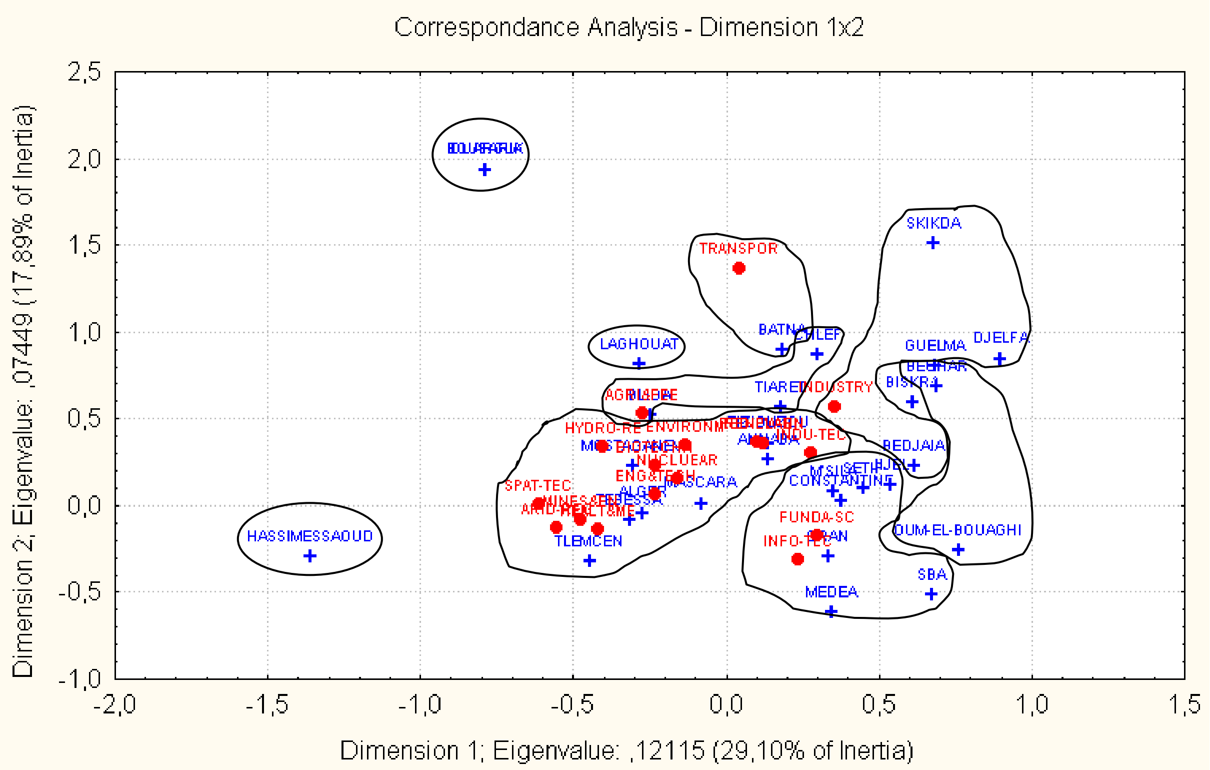


Figure 7: Transfer of the clustering results on the correspondence analysis map

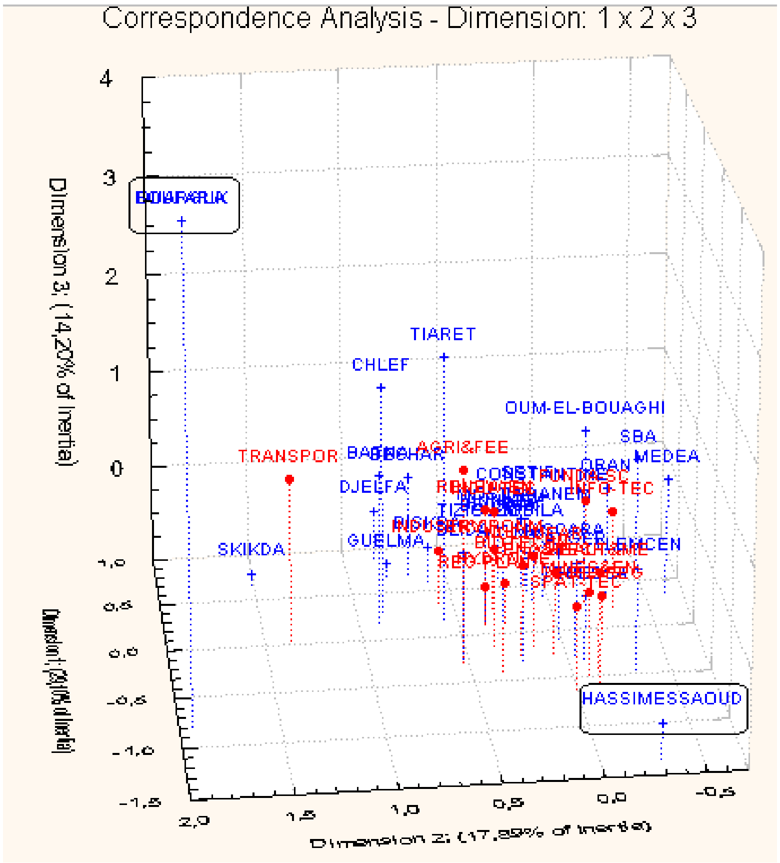


Figure 8: The correspondence analysis map with 3 dimensions

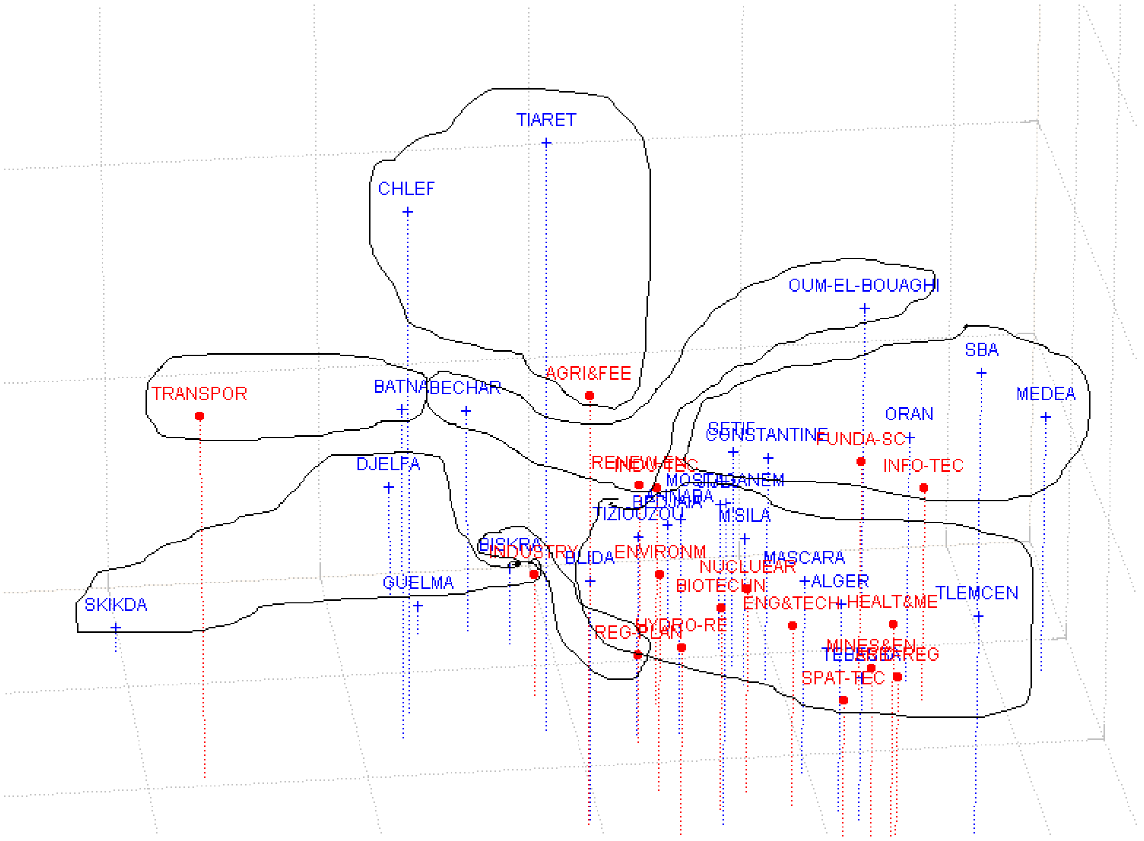


Figure 9: Magnifying of the core of the correspondence analysis map with clustering result transfer

Network drawing

The bibliometric mappings that implement methods of statistical data analysis (that we have just seen) are not always very easy to use. Their implementation requires good competences in these methods and especially of good experimental practices. This is why another type of graphical representation is often used for bibliometric studies: the networks analysis.

This method of representation has the advantage of being much more intuitive during interpretation because it is not based on complicated mathematical calculations. A network analysis represents visually the raw values contained in a co-occurrence matrix without additional mathematical calculation. The analysis network is more particularly suitable to the analysis of intra-field co-occurrence matrix (co-occurrence matrix of elements belonging to the same bibliographic field).

The network analysis is very used to map the networks of collaborations between authors (Fig 10) or inventors, between organizations (Fig 11) or companies, or for co-words analysis (keywords or topic network) or for co-classification analysis (IPC codes or database codes network).

Studies described the application of the analysis network in bibliometrics (Quoniam et al, 1995), web analysis (Rostaing, 2001) and competitive intelligence (Paoli et al, 2003).

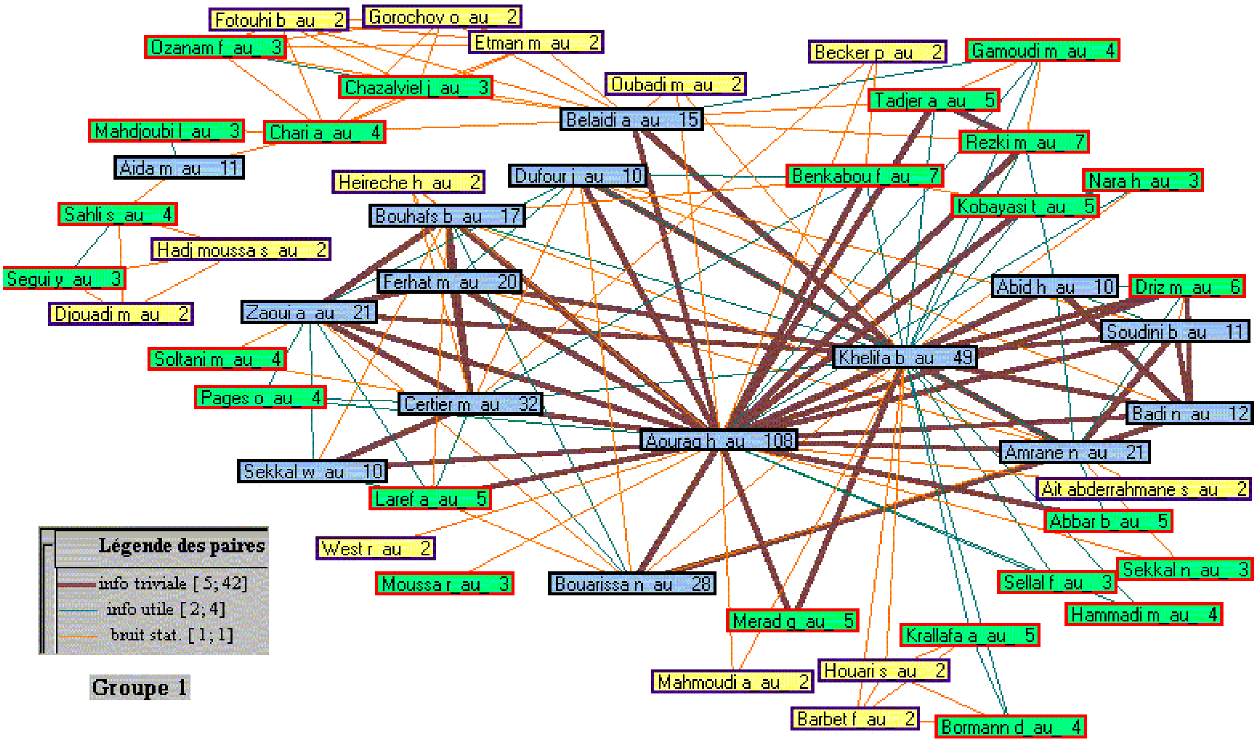


Fig 10: A collaboration network of authors publishing with Algerian scientists in Physics

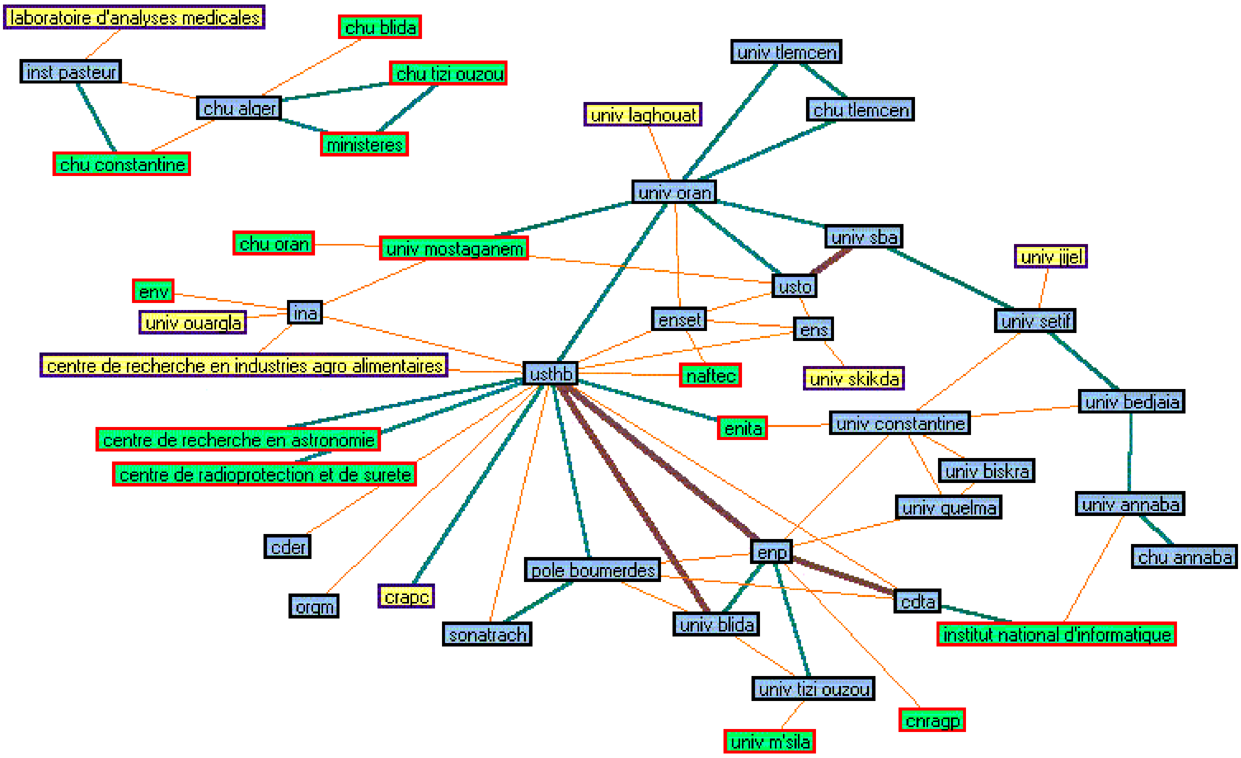


Fig 11: Network of National collaborations between Algerian institutes

Conclusion

Bibliometrics is especially suitable for supporting a Competitive Technical Intelligence process for companies. It is a valuable tool for assessing research and technology for the purpose of change forecasting and strategic development. These techniques help to identify emerging and declining researches or technologies and how those researches and technologies tend to sustain and borrow from each other. Also Bibliometrics is an efficient tool to allow identification of experts, centers of excellence, seminal articles and patents, and cans show trends of research and technology change. Bibliometrics is a decisive weapon for competitive intelligence.

Bibliography

Bisson C (2003), *Application de méthodes et mise en place d'outils d'intelligence compétitive au sein d'une PME de haute technologie*, Thesis: University of Aix-Marseille III

Bradford S C (1948), *Documentation*, Crosby Lockwood & Son, London, 156 p.

Catapano E (2001), *Conception d'un système de veille stratégique pour la détection systématique d'opportunités de développements technologiques et d'innovations : Applications aux PME de médicaments génériques*, Thesis: University of Aix-Marseille III

Da Silva A (2002), *L'information et l'entreprise : des savoirs à partager et à capitaliser*, Thesis: University of Aix-Marseille III

Dumas S (1994), *Développement d'un système de veille stratégique dans un centre technique*, Thesis: University of Aix-Marseille III

Escorsa Castells P, Rodriguez Slavador M, Maspons Bosch R (2000), “Technology mapping, business strategy and market opportunities”, Competitive intelligence review, Vol.11, N°1, p.46-57

Garfield E (1979), *Citation Indexing - its Theory and Application in Science, Technology, and Humanities,* John Willey & sons, New York, 274 p.

Kostoff R (1997), “Database tomography for technical intelligence: analysis of the research impact assessment literature”, Competitive intelligence review, Vol.8, N°2, p.63-79

Lauri P (1998), *Conception et gestion d'une cellule de veille technologique. Méthodologie et matérialisation d'un système d'information*, Thesis: University of Aix-Marseille III

Liège C (2003), *Conception et Mise en Place d’Outils de Traitement de l’Information Scientifique dans le cadre d’une Veille Sécurité Alimentaire*, Thesis: University of Aix-Marseille III

Lotka A J (1926), “The frequency distribution of scientific productivity”, *Journal of the Washington academy of sciences*, Vol 16, N° 12, p. 317-323

Nivol W (1993), *Système de surveillance systématique pour le management stratégique de l'entreprise. Le traitement de l'information brevet, de l'information documentaire à l'information stratégique*, Thesis: University of Aix-Marseille III

Paoli C, Dou H, Dou J-M, Mannina B (2003), “La construction d'indicateurs brevets par domaines technologiques” , *Cahier de la documentation*, N°2, p.45-59

Price D (1963), *Little Science, big Science*, Columbia, New York, 118 p.

Quoniam L, Rostaing H, Boutin E, Dou H (1995), “Treating bibliometric indicators with caution: their dependance on the source database.”, *Research Evaluation*, Vol. 5, N°3, p. 177-181

Rostaing H (2001), « Le Web et ses outils d'orientation. Comment mieux appréhender l'information disponible sur l'Internet par l'analyse des citations ? », *Bulletin des biblitohèques de France*, Vol. 1, p. 68-77, [url: http://www.enssib.fr/bbf/bbf-2001-1/10-rostaing.pdf]

Rostaing H, Léveillé V, Yacine B (2001), “Bibliometric study as an objective picture of the Algerian scientific research practices”, Proceedings: *The 8th International Conference on Scientometrics and Informetrics*, The University of New South Wales, Sydney, Australia, 15-20 July, p. 607-618

Small H (1973), “Co-quotation in the Scientific Literature: new Measure of the Relationship has between two Documents”, *Newspaper of the American Society for Information Science*, Flight 24, N°4, p. 265-269

Zipf G K (1949), *Human behaviour and the principle of least effort*, Editions Addison Wesley