Available online at www.sciencedirect.com

SCIENCE DIRECT®

Information Processing and Management 42 (2006) 1513–1531

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences

Michel Zitt [a,b,*], Elise Bassecoulard [a]

[a] *Lereco, INRA, BP 71627, F-44316 Nantes Cedex 3, France*

[b] *Observatoire des Sciences et des Techniques (OST), 93 rue de Vaugirard, F-75015 Paris, France*

Received 16 March 2006; accepted 16 March 2006

Abstract

Relevance of bibliometric indicators on scientific areas critically depends on the quality of their delineation. Macro-level studies, often based on a selected list of journals, accept a high degree of fuzziness. Micro-level studies rely on sets of individual articles in order to reduce noise and enhance precision of retrieval. The most usual information retrieval process is based on lexical queries with various levels of sophistication. In the experiment on Nanosciences reported here, this process was used as a first step, to delineate a 'seed' of literature. It has strong limitations, especially for emerging or transversal fields. In a second step, the alternative approach of citation linkages, was used to expand the bibliography starting from lexical seed. The extension process presented is ruled by three parameters, two deal with the cited side (threshold on citation score, and specificity towards the field), one with the citing side (threshold on the number of relevant references) interplaying in the 'referencing structure' function (RSF) introduced in a previous work. This type of combination proves effective for delineating the transversal field of Nanosciences. Further improvements of the method are discussed.

Keywords: Information retrieval; Lexical query; Citation network; Nanosciences; Scientific area delineation; Bibliometrics

1. Introduction

The delineation of scientific and technological fields is crucial to a series of decision-support studies: evaluation of institutions or countries, strategic positioning of actors, understanding of science and technology dynamics, based partially on the exploitation of publication and patents statistics and the analysis of S&T networks. Numerous commissioned studies, especially in Europe and in the US, address emerging or complex fields where the delineation issue is particularly difficult. This is the case of Nanosciences, which is taken as an example here.

[*] Corresponding author. Address: Observatoire des Sciences et des Techniques (OST), 93 rue de Vaugirard, F-75015 Paris, France. Tel.: +33 1 42 22 30 30; fax: +33 2 40 67 50 05.

E-mail addresses: zitt@nantes.inra.fr (M. Zitt), bassecou@nantes.inra.fr (E. Bassecoulard).

0306-4573/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2006.03.016

Several reasons account for these difficulties. The first one is the mere diversity of most of these large fields, which calls for multiple expertise to encompass sub-areas and translation in complex combinations of IR queries to cover the whole landscape. A second issue is the quick change occurring in emerging fields, hence in their scientific vocabulary and institutional structure, that necessitate frequent rounds of expertise if up-to-date queries are required.

Hybrid methods are more efficient to deal with such cases. Several types of bibliometric networks are associated to publication activity and all can be considered in delineation protocols. Here we limit ourselves to a two-steps combination of lexical and citation methods, but networks of authors and institutions could be helpful as well.

Section 2 is devoted to the Nanosciences context and general issues of field delineation, when using lexical entries or citation networks. In Section 3 we present the two-steps hybrid method. Section 4 reports a short characterization of outcomes. The last section is devoted to discussion.

2. Context

2.1. The context of Nanosciences

Nanoscience and technology is celebrated as a new horizon for science and industry, along with information-communication and biotechnology, and in relation with these sectors. The stakes of Nano both in industry and the military/security area, have fostered a series of national initiatives. A large body of literature in economics, management science and scientometrics, has already analysed the emergence of this area. Nanotechnology was one of the key areas selected in the EC 'cartography of excellence' exercise by FhG/CWTS (Noyons et al., 2003). Dynamics of the field combines bottom-up and top-down process (one of the dichotomies used by nano-experts reviewed in Fogelberg, 2003). Part of the literature devoted to the topic uses biotechnology as a benchmark, trying to point out their analogies and differences (Darby & Zucker, 2003). Laredo identifies 'Nano' as an example of new regimes of science (Bonaccorsi, 2002; Laredo, 2002).

The Prime program, a NoE at the EU level, entails a 'Nanodistrict' project aiming at the analysis of cognitive, industrial and territorial dynamics of this new area. The delineation of scientific literature on the one hand and of patent data on the other hand, are key stages for studying these aspects. Coping with the delineation issue by bibliometric methods is the object of this article. Bibliometrics is powerful as an arsenal of techniques to address scientific networks, but as a decision support technique it needs a framework or at least entry points set by scientists, experts or policy makers. A general and widely accepted definition of the field is an ideal entry point.

Does such a general definition exist for Nano? From its cradle in physics and engineering, nanoscience and technology have spread over a variety of disciplines, with the features of generic technology (e.g. Bachmann, 1998; NSF, 2002). More or less accepted definitions are in terms of scale range, such as Franks' definition of nanotechnology (Franks, 1987): *"technology in the range 0.1–100 nm (from the size of an atom to the wavelength of light) play a critical role"*. Some experts include the micron range (1000 nm). As technology and science in this area are closely related,[1] the extension of this definition to nanoscience may bring a large part of physics. Confronted with the issue of defining with acceptable precision what Nanoscience is and is not, for example by a positive list of subfields and their contents, experts hardly reach a consensus (Malsch, 1997, see also Glaenzel et al., 2003). The last authors report that the expert group appointed by the EC to prepare the NoE study on Nano only proposed the following "working definition": "*Nanotechnology – the manipulation, precision placement, measurement, modeling or manufacture of sub 100-nm scale matter*" (Meyer, Persson, & Power, 2001). The Europaeische Akademie (Schmid et al., 2003) chose, by contrast, not to mention a particular dimensional range in its "operationalizable definition of Nanotechnology" but to emphasize the specific

[1] Probably more than in any other field, technology is produced by scientists. Clear evidence is given by patent studies, which show that inventors are mostly scientists (Meyer, 2000; Thoma in the framework of the Prime Nanodistrict project, publications forthcoming).

size-dependent properties that "have no equivalent in the macroscopic world". "*Nanotechnology is dealing with functional systems based on the use of sub-units with specific size-dependent properties of the individual sub-units or of a system of those*".

Such definitions are not easily operationalized in bibliometric terms. A starting point of bibliometric delineation is the standard information retrieval method: lexical queries on databases using lists of terms elaborated by expert groups. Clearly the direct translation of the above definition would be inefficient, the combination of 'measurement' and 'nanometer' for example retrieves a lot of physics literature that would be ruled out from Nanosciences by most experts. It follows that bibliometric studies of Nanosciences had to rely on more or less extensive lists of descriptors trying to capture the variety of the field. Examples can be found in the above-mentioned works by Glaenzel et al. or Meyer et al., Noyons et al., following the pioneer work by Braun using simple entries (Braun, Schubert, & Zsindely, 1997). An elaborate formula has been set up in the final implementation of the EC 'mapping of excellence' by CWTS and Fraunhofer ISI was based on a list of keywords provided by experts of the field, commissioned by the EC. The authors also relied on classic additional look-up of journal sets and classification systems (Noyons et al., op. cit.).

2.2. Delineation and IR issues

The question of delineation is a particular case of information retrieval application where the object of a query is the selection of documents relevant to vast areas of scientific activity. Methods used in field delineation belonging to the standard arsenal of informetric methods, and the common framework of informetric distributions on the hand, IR performance apparatus, combining precision and recall, on the other, do apply. However, in comparison with run-of-the-mill IR tasks, some differences occur:

- The context of such studies is generally sensitive and require high standards of quality. This is particularly the case where science policy decisions are at stake. Conversely, the format of such studies usually allows sophisticated bibliometric means and escapes some immediacy constraints. For example, if some final data analysis process is used to reduce noise, the first stages of the study can focus on reducing silences, and redundancy of queries is less penalizing than in other contexts.
- The area encompassed may be broad and diverse. The organization of scientific areas reflects their growth patterns, with buds stemming from existing specialties. Most of basic scientific networks encountered (authors/institutions, citations, words) exhibit the power law distributions of the Lotka/Bradford/Zipf trilogy (for an overview see Bookstein, 1990a; Egghe & Rousseau, 1990; Rousseau, 1990) consistent with a variety of theoretical backgrounds (Bookstein, 1990b). In some interpretations, self-organization mechanisms and growth patterns tend to design embedded structures with some degree of self-similarity (Egghe, 2005; Katz, 1999; van Raan, 2000), overlap and fuzzy borders. The complexity of borders is a real challenge.

Commissioners generally put forth a broad definition of the field, possibly in political terms as well as technical ones. The problem is to translate this entry into a set of publications, by the mediation of librarians or bibliometricians. Information specialists may use the broad technico-political entries to trigger iterative IR search. Assuming that the delineation is made on a homogenous database, for example an ISI source (Science Citation Index, Web of Science, etc.), the issue is to find adequate protocols to:

(a) Make the best use of the complementary bibliometric networks by searching for
  - specific terminology;
  - key documents: specialized journals, review articles, etc.;
  - key actors (authors/institutions) on the topic (output, citations).

(b) Arrange the interplay of this process and experts' advice, possibly through iteration/learning.

Delineation protocols link these operations in various ways. Data-mining facilities allow a great flexibility for network exploring and iterations, including manual exploration and interactive selection, but at the same time more systematic processes embodying classical IR mechanisms are helpful to finalize delineation on complex fields. In this particular study, we examine a particular combination of lexical and citation steps.

- First step: the study starts from lexical queries built with the help of experts and published in the literature, that were combined and marginally modified; the final query adds the contents of specialized journals. This leads to a first bibliography on the field, used as a seed for the following stage.
- Second step: this bibliography is enriched by examining articles cited by the seed, the "cited core", and adding, to the seed articles, those who cite this core: the assumption is that these new articles share their intellectual basis (the cited core) with the seed. The aim of this 2nd step is to reduce silence.
- Third step: not reviewed here, it consists in reducing noise that could occur in the two first steps. It typically involves clustering lexical or citation networks on the extended set, in order to study border areas, detect topic that are too marginals or irrelevant, due for example to homonymye.

2.3. A framework for studying general retrieval conditions: the "referencing structure" function

It is convenient to turn towards a more disaggregated view to study the consequences of distributional features in a field on bibliometric analysis – whatever the object, words or citations. A blueprint of disaggregate analysis is the "referencing structure" function (RSF), first introduced for citations (Zitt & Bassecoulard, 1996; Zitt, Ramanana-Rahary, & Bassecoulard, 2003). In its original definition, this function describes in a closed field – assuming the literature known with a previous delineation – the fraction of this literature which can be retrieved under two interplaying constraints: a minimum threshold on citation scores for the cited repertoire Y, and a minimum closeness of the article with this repertoire, measured by the number X of references in common with this repertoire. Hence, in a field defined by a closed set A of citing articles referring to a set $B(A)$ of cited articles, the 'referencing structure' function is written $Z(X, Y)$. Z is the cardinal of the set of retrieved articles with $x \geqslant X$ references to articles cited $y \geqslant Y$ times. In a rank approach variant, the rank of citation score, instead of Y, is considered. This function entails citation and reference distributions that describe the field in a fairly complete way from the citationist point of view. It can be extended to word distributions, with some precautions (ibid.).

An interpretation is in terms of a simplified graph of citation. Since only first-order citations are considered, a simple representation is a couple of sets A and B, respectively, emitting and receiving citations (Fig. 1). For simplicity sake, A and B do not overlap, a condition that can be easily relaxed. The function $Z(X, Y)$ is a measure of the robustness of the graph in a particular sense: it represents the number (or proportion) of nodes in A connected with a node degree $\geqslant X$ (at least X links to B), when lower degree nodes of B (degree $< Y$) and their attached edges disappear.

Cuts of the RSF (for example Z for Y fixed) can be approximated by Weibull laws (or related forms):

$$Z(X, Y = y)/Z(1, y) = \exp\left(-\left(\frac{X-\xi}{\alpha}\right)^c\right) \quad X, Y \geqslant 1$$

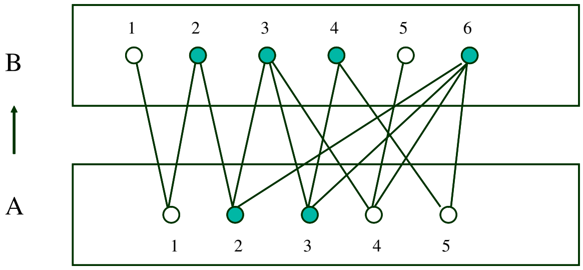


Fig. 1. Bipartite graph: illustration of the "referencing structure" function (RSF). $Z(X, Y)$ is the number of nodes A (citing) remaining with a degree $\geqslant X$ when nodes B (cited) with a degree $< Y$, and the attached edges, are deleted. For example, in the graph above $Z(3, 2) = 2$.

where the exponential term is the survival function of the Weibull distribution:

$$\mathrm{SW}_{(\xi,\alpha,c)}(X) = \exp\left(-\left(\frac{X-\xi}{\alpha}\right)^{c}\right) \quad \text{for } X > \xi$$

The particular case $c = 1$ gives the exponential distribution. In our previous tests, it corresponds to cases where the threshold y is not low.

Now let us make the assumption that the contribution of each article to the topic is proportional to the number of specific references X. It can be demonstrated (*forthcoming*) that in the case $c = 1$ the corresponding IPP (Information Production Process introduced by Egghe, 1990) is such as the concentration of information (each retrieved article contributing in proportion of its number of relevant references) in function of sources is

$$\mathrm{SR}_{\alpha,c,1}(X) = w\left(1 + \log\left(\frac{1}{w}\right)\right)$$

where $\mathrm{SR}(x)$ is the cumulation of items, and w the cumulation of sources. This corresponds to a concentrated scheme, however, less concentrated that the typical Bradford–Leimkuhler situation.

The RSF framework is useful in two occasions in this study: for the global comparison of retrieval based on words and citations; for the crude modeling of the extension process.

2.4. Power and limits of lexical delineation

Lexical retrieval seems powerful in quantitative terms. The RSF allows us to compare in a crude way the general retrieval properties of words and citations. Let us imagine that by some external device (previous studies, experts consultation) we have collected structuring items for a given field (say the n more frequent cited articles, and n more frequent not-void terms, respectively, corresponding to frequency thresholds Y1 and Y2). We wish to retrieve the corresponding literature that uses these words or cites these cited items. Then we put the simplest constraint of relevance, by demanding that, to be retrieved, an article should have at least X terms, or X references, in common with the selected set of structuring items based on Y1 or Y2 thresholds. It will typically appear that in the range of low n (high Y) the lexical recall is quantitatively more efficient. A trivial example is by choosing $n = 1$ and $X = 1$ the number of retrieved articles is equal to the frequency of the top word, almost always more frequent than the top-cited item. A sketch of the number of articles retrievable for the same n (Fig. 2, based on the same set, a large extract of our "Nano" file) shows the advantage of a lexical approach from this quantitative point of view: a lexical query based on short lists of most frequent (non-void) words (word-1 OR... word-*i* OR... word-*n*) retrieves a much higher number of articles than a similar query based on the n first cited items (cited-1 OR... cited-*i* OR... cited-*n*) – a rather general result. The lexical query is in this respect more efficient than a citation-based query of the same length, for sensible values of X.[2] The difference would be still stronger if the same Y threshold, instead of n, were applied to words and citations.

This theoretical advantage of lexical queries, in terms of retrieval (number of articles), should be checked in terms of information, i.e. the IPP, but even if similar hypotheses and calculations may be carried out on each side, the ground for a sound comparison is lacking: is an article retrieved by X references generally more relevant that an article retrieved by X words (or any sensible weighted formula)? Besides, the quantitative advantage of lexical queries is balanced by several difficulties.

Constraints on retrieval based on words are strong, as shown in the abundant literature in IR, linguistics and Natural Language Processing. ISI databases do not provide controlled terms as such. The difficulties of natural language (traps of terms extraction, homonymy, synonymy) require sophisticated treatment and, most of the time, human scrutiny for each term involved. General issues are well known. When addressing the bibliometric delineation of a field, problems of unification and disambiguation may not be more severe than in

[2] The inversion of positions for high values is trivially due to the length of references lists vs. title and keywords lists.

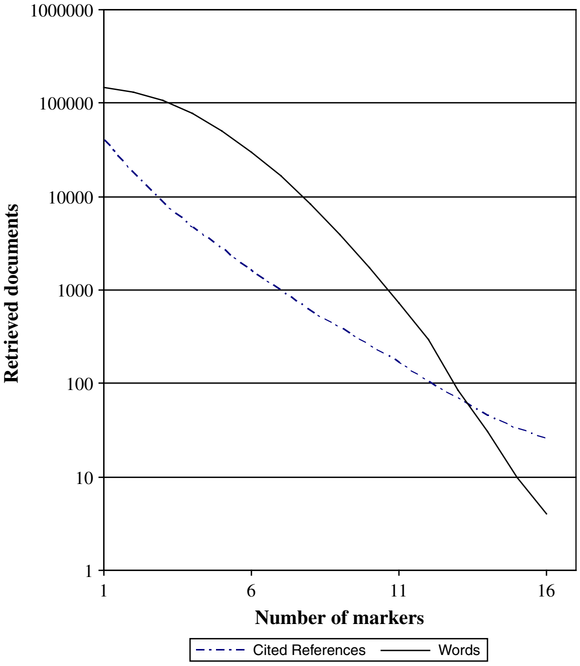


Fig. 2. Number of retrieved documents, for the $n = 200$ most frequent cited references, respectively, the most frequent words, as a function of the number of markers X, in the experimental set.

usual information retrieval applications, but the scale adds to the difficulty. For example, a major source of noise, homonyms, tend to be more prevalent, and generally more difficult to disambiguate, than in small contexts. The specification of all synonyms is also a difficult task, requiring previous text processing, for example lexical clustering of the field vocabulary on the basis of co-occurrence indexes.[3] Acronyms are a particular kind of synonyms. For example, Biotechnology or Nanoscience fields make use of hundreds of acronyms, especially for techniques and methods. Querying on these particular synonyms enhances recall, but as they are unambiguous only in narrow contexts, the risk of homonymy explosion may deter from using short acronyms.[4] The acronyms are a good case of IR trade-off.

At the field level, markers both synthetic (frequent in the field: not silent) and specific (not frequent elsewhere: not noisy) are particularly helpful. For example, publications in Nanosciences are expected to use frequently terms with the prefix NANO. A truncation using such a prefix will alone retrieve a large number of articles. Noise associated to the prefix NANO is manageable, this prefix is reasonably specific.[5] But on the border of Nanos, some articles rather use an entry through MICRO scale (sometimes termed the top-down or miniaturization approach in Nanoscience and Nanotechnology). Clearly, the absence of specificity of the prefix MICRO toward the field will lead to a heavy process of specification either by combination (AND clauses in a Boolean scheme) or exclusion (NOT clauses). Many similar examples could be found in attempts of delineation of transversal fields, like climate, environment, etc.

Field-level markers cannot encompass the variety of the field. Identification of sub-areas (whatever their nature: techniques, products, materials, processes, etc.) is necessary to reduce silence. Basically, if we refer

[3] To a certain extent, this process has already been used for improving the lexical formulas used here. For example, co-occurrences with ‘Nano’ terms, however, not systematically, were investigated.

[4] Alleviated in ISI sources by the use of authors’ keywords, which often associate the acronym and the complete form.

[5] Trivial noise with chemical forms in NaNO; with units of measure: nanometers, nanosecond, nanojoule, etc. – but some ‘nanometers’ are relevant; with some living forms (nanoplankton).

to the embedded structure of networks, the problem of delineation is merely transposed at a lower level,[6] and multiplied to cover the whole field by the mosaic of (overlapping) sub-areas. The lexical query used here is to a certain extent such a mosaic, expressed by the sequences of OR clauses. An obvious limit of the process is the fact that recording *ex ante* the list of sub-areas and of their specific vocabulary is hardly workable.

The lexical approach is quite difficult to implement without the active participation or supervision of experts aware of the vocabulary of fields and sub-fields. Even in the basic task of building anti-dictionaries of too broad terms, expert intervention is necessary. However, as soon as they are committed to subfield descriptions, the risk of specialization arises. Updating studies based on lexical delineation is not straightforward: while vocabulary should be defined periodically because of relatively swift changes, the intellectual base, at least its highly cited core, can be updated in more automatic ways. The theoretical advantage of the lexical approach mentioned above on the RSF is partly misleading since a list of words needs almost always some supervision, while citation processing can be largely automatic.

In addition to the informetric features recalled above, this designates a lexical approach for the first-step of the delineation process (the seed), which can also include some other queries less sensitive to local traps, for example, journals expected to address the entire field (in our example, journals bearing "nano" in their title). Citation methods intervene in the second step.

2.5. Citation methods

2.5.1. Formal analogies and contrasted properties

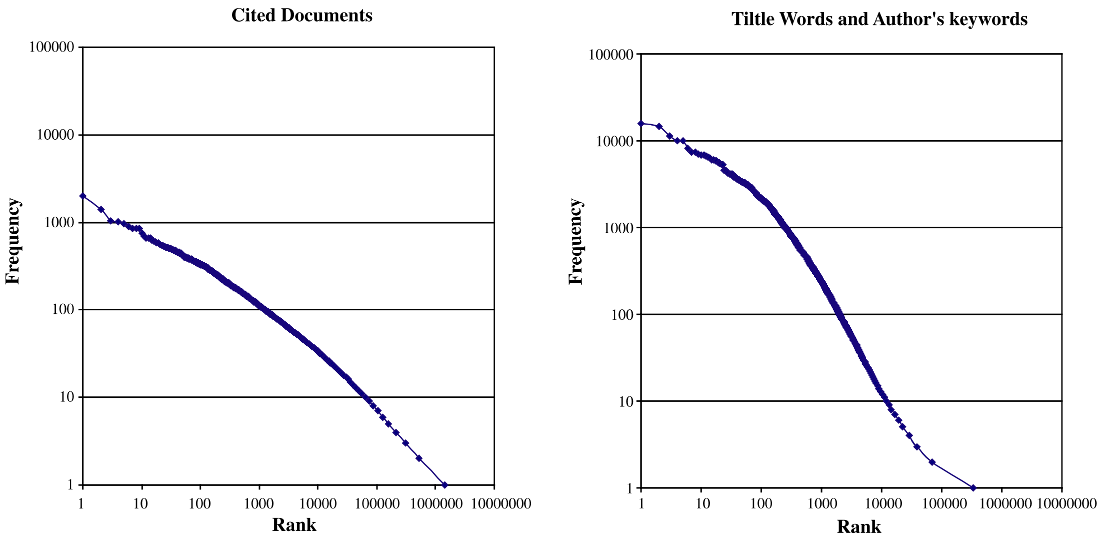
The mainstream of information retrieval has been based on lexical applications, and most classical models in the domain have used term-based queries. In the sixties, an alternative or complementary way appeared for scientific information, with ground-breaking works of Garfield on citation indexing (Garfield, 1967). Although other networks (co-authoring for example) may be used for structuring purposes, the lexical and the citation approaches appear as the two main ways of investigation, either competing of complementary, in a large class of IR or bibliometric issues.

Statistical features are similar enough, with distributions close to power laws, with tail irregularities. The literature devoted to linguistic distributions in the wake of Zipf is abundant, and several models on citation distributions can also be found since Price (Naranan, 1970; van Raan, 2001; extension to hyperlinks for example by Egghe, 2000; Rousseau, 1997). In citation indexing, bibliographic references can be used as an extended form of lexical tokens, likely to undergo similar forms of querying. The parallelism between the two forms of indexing is deep enough to allow the transfer of methods and algorithms: analogy of lexical coupling of documents and citationist 'bibliographic coupling' (Kessler, 1963), formal similarity of co-word (Callon, Courtial, Turner, & Bauin, 1983) and co-citation (Marshakova, 1973; Small, 1973).

Both word and citation distributions lie in the range of strong concentration. However, parameters are clearly different. Word distributions are more concentrated. The Zipf–Mandelbrot exponent can be different from the canonic value of one, depending on the richness of the language, which also depends on the type of language (natural, controlled). Rich language implies slightly less concentrated distributions. Citation distributions tend to be less concentrated, with a dependence on dynamic features, namely the citation time-window used. A classic interpretation of this difference in exponents is that the citation network exhibits a higher degree of complexity (or is more fractal) than the word network. Fig. 3 gives an idea of the two empirical distributions on a same universe, the final "extended" set described below. The distribution considered on the lexical side is the distribution of title words plus authors' keywords.

Beyond the formal analogy, interpretations of networks differ. The differences in nature are well known, and highlighted in the citation studies literature following Garfield on citation indexing (Garfield, op. cit.). Citation linkages are primarily diachronic and unidirectional, a feature that remains exploitable even in further transformation into symmetrical linkages (co-citation or bibliographic coupling). In the canonic presentation of cocitation research fronts (Small & Griffith, 1974), co-cited cores describe the intellectual basis of

[6] Moreover, an optimum level of decomposition of scientific networks is hardly found (e.g. Zitt, 2005).


Cited Documents

Tiltle Words and Author's keywords

Fig. 3. Frequency–rank distribution, for citations and vocabulary (title words + authors' keywords). Vocabulary has been only partly cleaned (plural forms, English/American; no unification of acronyms at this stage). Ranks: low tie option. Based on a partial processing of citations (>90%) in the experimental set.

current activity, the research fronts as such, which dynamically build on combinations of former research. Citations may also reflect directly sociological linkages amongst scientists, for example schools of thought.

As lexical querying has been challenged by citation indexing, conversely structuring of scientific fields by citations has been challenged by co-word studies (Callon et al., op. cit.). The same is true for relations between science and technology where citation networks can be challenged by lexical linkages (Bassecoulard & Zitt, 2004).

As a result of this partial analogy, the outcomes of IR or bibliometric trials using the two methods converge only partially. In a comparison of searching by citation or by lexical approach (Pao, 1993), Pao advocated using combinations of methods. In practice, the two approaches are often used sequentially. The same is true for thematic analyses based on co-word and co-citation, for example.

2.5.2. Complementarity

In bibliometric applications, citation-based techniques cannot spare a minimum of lexical analysis, for example for giving titles to co-citation fronts. More elaborated forms of hybridization have been practiced. ISI 'keyword-plus' also result from a lexical elaboration on citation-related articles. Other examples are hybrid methods are found in Braam, Moed, and van Raan (1991) or Leydesdorff (2004).

If we focus on IR, the classical way of lexical queries is still dominant. However, due to the widespread diffusion of the Web of Science and citation engines such as CiteSeer, the retrieval techniques based on citation indexing are developing. Since Garfield's early works, bibliometricians have used citation linkages to delineate scientific areas at coarse-grain level (journal) or at fine-grain level (individual publications). Data-mining techniques have given a new momentum to this approach (Kostoff, delRio, Humenik, Garcia, & Ramirez, 2001). Other bibliometric networks (authors, institutions) may also be mobilised. Hybrid combinations are often specific to particular studies.

One hybrid combination is famous world-wide: exploiting the analogy of hyperlinks and citation linkages, the Google search engine implements in principle a combination of lexical query used as a seed, and an adjustment by an iterative qualification of hyperlinks nodes (Brin & Page, 1998), leading to the 'Page rank' as a probabilistic relevance measure of the page based on random web surfing following the citation links. The context of bibliometric citations is somewhat different, with unidirectional linkages pointing to the past, except in rare cases of quasi-simultaneous works.

Handling citation networks in an initial round would not have been straightforward. Experts can spontaneously describe a field in terms of words. They might be asked instead to post some key publications or authors, but they would encounter difficulties to enter long lists of those. Some length is required to compensate the less favorable retrieval conditions due to a lower concentration in citation distribution. Fortunately, these lists can be easily generated by automatic means, as soon as a seed is already available from a first step, and need very little supervision. Some technical equivalents of synonymy and homonymy do exist in the matching processes necessary to handle citations,[7] but these issues are purely technical and less severe than for words. The combination of a first step using the lexical/expert way, and a second step based on citation relations, seems sensible.

3. A two-step hybrid lexical-citation method

3.1. The nano field described by terminology

The Nanoscience field exhibits several peculiarities that hinder or favor the quality of the delineation by lexical means.

- Multidisciplinary character: Nanoscience and technology as a generic toolbox irrigates a large variety of fields: electronics and communication, biotechnology, materials, physics, etc.
- Emergence: the field is still rapidly growing and new buds can appear without stabilized vocabulary.
- 'Brandname' effect: the term or prefix 'Nano' seems to act as a label likely to attract attention from scientific community, industry and funding bodies. Authors tend to post it as a beacon in titles or author's keywords. This posting process seems explosive (Schummer, 2004) and, so far, stronger than the reverse behavior, the avoidance of the term, which might emerge from the perceived threat of Nanotechnology by some fractions of the public or NGOs. As long as 'Nano' is used as a label, the lexical delineation is made easier, and this field appears among the 'best cases' for a lexical approach. However, the analysis of the extended set shows that the 'Nano' token is missing in many relevant articles.
- Previous experiments. The lexical formulas recently published, properly combined and adjusted, appeared as a good starting point to build the 'seed'.

The core of the formula was established after the EC study above-mentioned, which presented two distinct queries, science-oriented and technology-oriented, established, respectively, by CWTS and FHG-ISI. It appeared that the technology-oriented formula, adapted by Thoma at SSSUP,[8] also retrieved a relevant set of articles in the ISI database, in addition to those recalled by the science-oriented query. We eventually combined the two queries. We added input from Meyer et al. preparatory study (op. cit.) and our own modifications and adaptations, based on co-occurrence checking. The final query is presented in Appendix.

The strategy of this formula is to capture many aspects of Nanosciences without taking the risk of a noisy explosion: acronyms are avoided; restriction by combination is used, including for topics historically close to Nanosciences, such as fullerenes; several sources of noise (nanometer for example) are specifically treated. One can expect that the main risk associated with this set of queries is a relatively high level of silence. The query was applied to ISI database to create the 'seed set' A, in a Boolean logic. Due to limitations of software, no weighting was used. For example, the internal frequency of a term in an abstract does not count, nor the number of sub-queries hit by an article. The outcome is a non-fuzzy set without ranking.

3.2. Enrichment by citation analysis: principle

The main objective of the enrichment is to reduce silences unavoidable in the lexical query. Basically, we try to enrich the initial seed by articles exhibiting the same intellectual basis as the seed, that is, if we use a

[7] ISI matching were used for this particular extension process. Some technicalities of matching were studied by Moed and Vriens (1989).

[8] Co-workers in the Prime nano-district project.

Mertonian interpretation, the articles citing the same cited articles as the seed. The process relies loosely on a 'bibliographic coupling' rationale (Kessler, op. cit.), but instead of coupling for pairs of individual articles, the similarity is considered at the level of the whole sets, by the following operations:

1. Building the set B of items cited by articles in the set A (the 'seed'), with their score of citation y (citations received from A).
2. Building C, subset of B by thresholding on y. Only articles such as $y \geqslant Y$ are kept. Y is the first parameter of the extension process.
3. Introducing a second threshold based on the ratio $u = y/y'$, where y' is the citation score calculated on the whole database S. Only articles with $u \geqslant U$ are kept, forming the set D (or 'core'). U is the second parameter of the extension process.
4. Building, on the citing side, the set E of citing articles with $x \geqslant X$ cited references belonging to the core D. X is the third parameter of the extension process.

The set (E minus A) represents the extension.

The sets A and E, partly overlapping, sharing the same repertoire of citations, are 'bibliographically coupled' at the set level. Except in limit cases, set-level coupling retrieves more articles than document-level coupling, because it allows to retrieve articles with combinations of cited items not existing in the references of individual A (see below). Another difference is practical: set-level coupling requires fewer computer resources, since no item-to-item proximity is calculated.

The set A may contain articles not present in E. In the present experiment, this is the case since A was built on lexical queries and contains articles with less than X references to the cited core with the parameters chosen.

The process could be iterated. The sequence is iterated starting from E or Union(A,E) as the new seed. In the present experiment, the enrichment was largely sufficient for a sensible setting of parameters below, and the process was not iterated.

Fig. 4 illustrates the principle of the protocol.

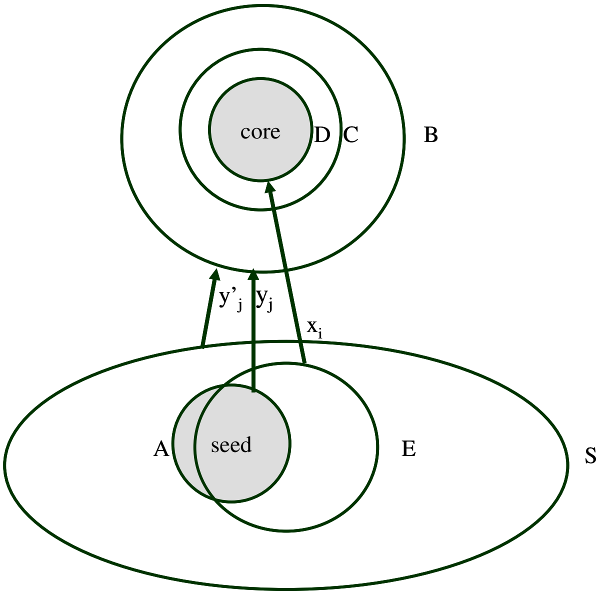
3.3. Interpretation of the parameters

The high connectivity of citation networks imposes strong restrictions on the extension process. In a non-constrained protocol, all articles in S citing a single article in B belong to the extended set. The cardinal of this set is very large, and so the amount of expected noise. Each of the three parameters needs some interpretation.

The citation score threshold Y. Articles with few citations are less likely to have a structuring effect on the universe, moreover they are likely to focus on particular subfields – possibly artefacts due to experts' specialization – rather than the common intellectual basis of the whole field. On the other hand, keeping Y low helps to reduce silence on the neighborhood of the original set. Applying a threshold has also practical aims: keeping long lists of cited references, in such a skew distribution, is costly for value. Y may be termed a '*genericness*' parameter.

However, the set C contains many articles which are not at all specific of the field. For example, in the Nano seed, many articles of general physics or fundamental biology are cited. If we keep these cited items, a large number of citing articles will be retrieved that have little or no relation with Nano field. A specificity threshold U was calculated for cited articles, ensuring that the ratio u of the local score y to the global citation score y' of each cited article in D is large enough.[9] Some implementation of specificity ratios would be necessary as well for extension processes based on vocabulary (Noyons, 1999). The setting of U is somewhat tricky, a trade-off should be sought between the capability to get out of the local traps suspected in the lexical query, capability which grows inversely with U, and the noise which decreases with U. U may be termed a '*specificity*' parameter.

[9] u and U might be normalized by a ratio of size of A and S for a probabilistic interpretation.



CITED SIDE

Set B all art. cited by A

j(1..m) cited articles

Set C = B with y_j>=Y

Set D (core) = C with

u=y_j/y'_j>=U

FLOWS

y_j # cit from seed to j

y'_j # cit from database to j

x_i # ref in article i to core

CITING SIDE

Set A seed:

i(1..n) citing articles

Set S whole database

Set E: all art.

with x>=X ref to D

E-A: extension sensu stricto

Fig. 4. The extension process. The citing seed *A* yields the cited set *B*. Applying the threshold *Y*, *B* is reduced to *C*; applying the threshold *U*, *C* is reduced to *D*. *E* is the final set of articles citing *D* with the threshold *X* on their references.

Applying a constraint of referencing. The simplest constraint is setting a ‘referencing threshold’ *X* to the number of references that a citing article places in *D*, say *x*. The assumption is that the more references in the citation repertoire of the seed, the more chances of relevance towards this field. The idea that *X* is a proxy for relevance is strongly supported by an earlier study (Zitt & Bassecoulard, 1996). *X* and *Y* are substitutable for achieving a given level of retrieval *Z*, but in terms of precision the combination ‘high *X* and low *Y*’ perform better than the reverse. *X* will be termed a ‘*relevance*’ parameter. The “referencing structure” function RSF quoted above is helpful to describe the extension process, with some modification due to the openness of the sets (see Section 4). Considering *X* as a proxy of the informative content towards the field allows to study the corresponding Information Production Process mentioned above.

4. Application and results

4.1. The effect of parameters

An adapted form of the referencing structure function is helpful to describe the extension process. Rather than the RSF built on either closed set of citing articles, *A* and *E*, we consider an intermediary construction here, with *Z* calculated on *E*, on the citing side, and on *B*, on the cited side. A difference with the original RSF

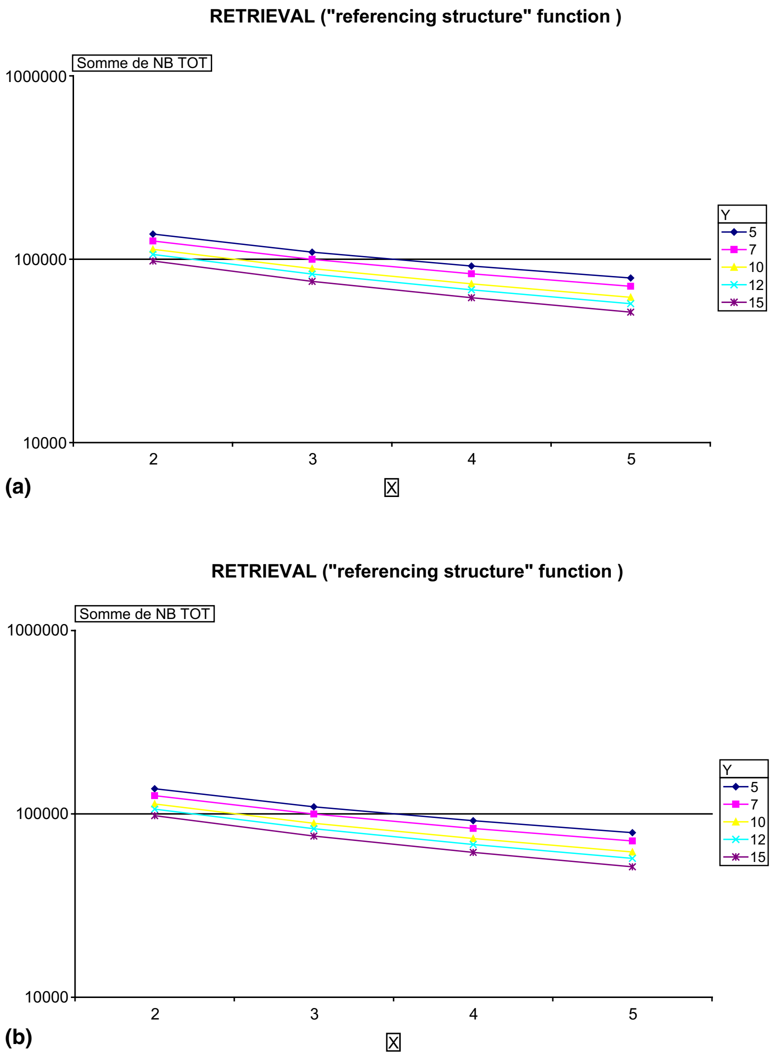


Fig. 5. Total number of retrieved documents for selected values of the parameters U (specificity), Y (genericness) and X (relevance). A: $U = 0.3$; B: $U = 0.5$.

comes from the fact that the threshold Y is complemented with the threshold U, capturing information from the whole database. All things being equal, Z grows inversely with $U[0+, 1]$, and is minimum for $U = 1$, corresponding to references strictly specific of the seed set A. For U minimum, the extension is too large to be manageable.

Fig. 5 shows the retrieval level for several combinations of the parameters, as a function of X. Ordinates represent the total citing literature retrieved, i.e. comprising the intersection with the seed. A fraction of the seed is not retrieved because articles do not meet the requirements expressed by the thresholds on parameters. Fig. 5 suggests that, if a Weibull distribution is used to model the retrieval, the characteristic exponent c is slightly less than one. For a given Y threshold, large changes in retrieval occur for the low values of X, for example when shifting from $X = 2$ to $X = 1$. Allowing low values of X leads to an explosion of literature, likely to carry a high level of noise.

For example, the triplet ($Y = 5$, $U = 0.3$, $X = 4$), would lead to an extension of almost one half of the seed, after some filtering on the seed (articles with references, presence of addresses, etc.):

Original set (seed): 122,000
Extension: +56,000
Final set: 178,000

This degree of extension would still be sensible. In the present implementation, a mixing of more restrictive strategies was used for building an experimental set, trying to address synonymy/equivalences on one hand (low *U*), and identifying 'new' topics on the other. It led to the addition of between 25% and 30% more publications to the seed, depending on the filters. The tables of contents shown here are based on this experimental set.

Contents of the 'extension': subfield diversity

Preliminary results suggest that the citation-based extension does not reproduce the structure of the seed. In the seed, four fields are prominent (material sciences-multidisciplinary, physics-applied, physics-condensed matters, chemistry-physical; in that order). They are also dominant in the extension, but in reverse order. Then chemistry-multidisciplinary, polymer science, physics-multidisciplinary, physics-atomic-molecular-chemical, chemistry-analytical are reinforced in the extension while engineering-electrical-electronic, crystallography and metallurgy are weaker. Globally, the two first fields lose some ground in the extension, but the 'next best' are reinforced. Further analysis at a more detailed level (journal level, research fronts) is required to validate the idea that citation-based extension reduces the specialization effects.

Fig. 6 shows the distribution amongst fields.

Contents of the 'extension': topics

We report here a first analysis of the contents of the extended set, in terms of vocabulary, limited to title words and authors' key-words, the latter not always being present in articles notices. For this preliminary task,

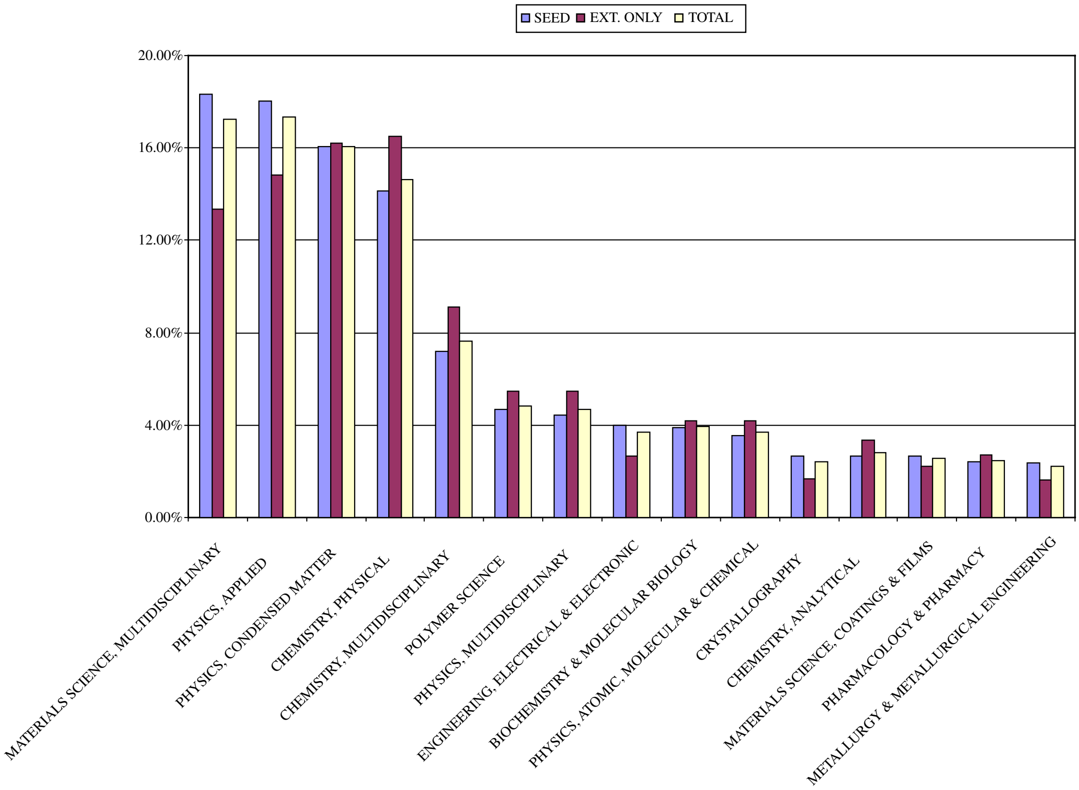


Fig. 6. Distribution of documents among subject categories: seed, extension only, total. Ordinates: % of documents in which the category appears (full count) on the experimental set.

we limited ourselves to simple natural language processing operations. To study the composition of the extension, we first calculate measures of relative frequency for all terms in the seed and the extension. Terms more present in the extension are designated as 'new' in the following. Some of the 'new' terms above, both frequent and specific to the extension, could be central to 'new' themes.

(a) If so, they are likely to come up with a context which is also new, i.e. typical of the extension. We selected a few candidates and examined whether their associated words were also specific to the extended set. A co-occurence index (Ochiai) was calculated for these terms on a subset of articles whose titles contain at least one multiterm or acronym with a global frequency $\geqslant 5$.

The extension brings topics absent from the initial seed either because the corresponding terms are not present and have no equivalents in the original query, or because terms were identified in the query, but associated to restrictions, e.g. AND or NOT clauses. An example here is '*self-assembly*' (Table 1A).

Several 'new' themes concern spectroscopy, micro and mesoporous materials (that may be included by the definition of Schmid, op. cit.) and various types of thin films. Preliminary results on clustering[10] both on words and citations confirm that mesoporous materials clusters, for example, show a high proportion of articles from the extension set. Another example (Table 1B).

The reasons for the absence are various: terms forgotten by experts; terms deliberately excluded because of the fear of noise generated by too general or ambiguous formulation; terms deliberately excluded as deemed out of the scope of the topic.

(b) If terms are new within an 'old' context, they may contribute to the enrichment of themes already present in the seed, rather than creating new themes. In some cases, new terms with any frequency are simply synonyms or acronyms of 'old' terms (Table 1C).

5. Discussion and conclusion

Lexical queries can hardly encompass all the aspects of a field. The specialization of experts, the difficulty of generating efficient queries (for example of adequate restrictions when using general terms) can lead to irregular delineation where one aspect of the field is correctly addressed while others are not properly dealt with. When one faces the issue of field delineation, a sensible aim is a balanced exploration of its various facets. The citation protocol is expected to get out of the constraints of lexical querying systems, on the one hand, and the traps of supervising experts' specialization on the other. The capability to do so is ruled by the setting of extension parameters, especially the specificity threshold U. If a high selectivity is set, the extension of the set will not go beyond the narrow neighborhood of the seed. The risk is to remain stuck in the close neighborhood of the original query. A lower selectivity threshold will allow to go further, but, all things being equal (the other parameters) at a risk of explosion of literature. The parameter X ruling the relevance of the extension is also crucial. Formal apparatus ('Referencing Structure' Function, Information Production Processes) help to describe the properties of the process. Applied to the field of Nanoscience and technology, the method proved efficient when using a prudent combination of parameters, aiming at a moderate addition to the seed and based on a rather secure combination of parameters: high specificity ratio, and no less than four references to the cited set. A first comparison of the contents of the extension and the seed was conducted.

Several aspects can be discussed:

Dependence on citation interpretation

The Mertonian terminology has been used in this text, for example the notion of "intellectual basis". However, the informetric process described supports different interpretations, for example if the cited cores are considered as "shared legitimatory repertoires" (Rip, 1988) or to some extent actor-network views, as soon as the sharing of cited references – whatever the nature of the linkage – is accepted as a sign of contextual proximity.

[10] Clustering tests are conducted in collaboration with Alain Lelu (LASELDI, Université de Franche-Comté). Complete results will be reported later.

Table 1A

Term	10 first neighbors
Self_Assembly 8 ‘new’ neighbors out of 10	Supramolecular_Chemistry Hydrogen_Bonds N_Ligands Molecular_Recognition Hydrogen_Bonding Helical_Structures Cage_Compounds Block_Copolymers Crystal_Engineering Scanning_Tunneling_Microscopy

Table 1B

Term	10 first neighbors
Molecular_Recognition 8 ‘new’ neighbors out of 10	Supramolecular_Chemistry Molecular_Tweezers Synthetic_Receptors Hydrogen_Bonds Host_Guest_Chemistry Molecular_Meccano Hydrogen_Bonding Molecular_Imprinting Host_Guest_Complexes Template_Directed_Synthesis

Table 1C

Term	10 first neighbors
Quantum_Wells 9 ‘old’ neighbors out of 10	Molecular_Beam_Epitaxy GaInNAs_GaAs Semiconducting_III_V_Materials Optical_Properties Structures_Grown Laser_Diodes Electronic_States_(localized) InGaAs_AlAsSb Long_Wavelengths Quantum_Dots

Variants

Some variants are worth examining, without getting out of this 3-parameters protocol. On the citing side (x):

– We used an absolute measure, the gross number of references x, but a weighting by the number of references in each citing article, could be envisaged. That citation habits differ across scientific communities is well known (Murugesan & Moravcsik, 1978). The difference may be considerable for designing knowledge networks, especially when citing articles belong to several disciplines where the average lengths of the reference lists are different with severe consequences for various comparisons (Zitt, Ramanana-Rahary, & Bassecoulard, 2005). Metrics with fractional measures were used in co-citation contexts by Small and Sweeney (1985) and among others by Zitt and Bassecoulard (1996).

- Another weighting scheme may use a frequency weight (based on the score y) instead of the gross count for each of the x references in a given citing article. A classical scheme uses TF-IDF score (Salton & McGill, 1983). Other schemes avoid an over-rating of very low frequencies, likely to carry errors.[11]
- On the cited side, an iterative qualification of cited items is possible following Pinski and Narin's proposal ('influence' measure as a refinement of Garfield's "impact" measure, Pinski & Narin, 1976). Scores of cited articles are weighted by the score of citers. These measures are more adapted to lasting entities (journals, authors perhaps) than to dated documents, where the management of citation delays could be difficult.
- For memory sake, a lexical extension may be conducted rather than citationist extension, but no advantage is taken of the complementarity of methods.

Further analyses

The comparison between seed and extension will be pursued after a mapping exercise, where the 'new' themes are expected to appear as clusters. The capability of citationist extension to 'smooth' the delineation will be tested. Another question is the setting of parameters for such extensions: in absence of theoretical optimum, rules of thumbs could be explored through new experiments.

Acknowledgements

The authors thanks Thierry Lafouge (University Lyon-1), Suzy Ramanana-Rahary (OST), Pierre-André Zitt (PhD student, University of Nanterre), Alain Lelu (LASELDI, University of Franche-Comté) for helpful discussions; Eric Vachon and Nelson Teixeira (OST) for developing the software.

This work is part of the Nanodistrict Project (V. Mangematin, UMR Gael, INRA, coord.) within the NoE Prime (P. Laredo, LATTS, ESPC, coord.).

Appendix. Initial query (adapted)

All queries on textual fields (including abstracts) of notices, query 1 also includes NANO in journal titles. (*) denotes the right truncation.

Query 1: NANO*:

Articles excluded: where NANO occurrence is only due to forms of (NANOMETER, NANOLITER, NANOAMPERE, NANOCURIE, NANOJOULE, NANOKELVIN, NANOTESLA, NANOWATT, NANOSECOND, NANOGRAM) or NANOMOLE-NANOMOLAR *or chemical formulas* NaNo*.

Query 2a: (NANOMET*-CHIP* OR NANOMET*-LAYER* OR NANOMET*-DIAMET* OR NANOMET*-ELECTRON* OR NM-ENGIN* OR NM-CHIP* OR NM-LAYER* OR NM-DIAMET* OR NM-ELECTRON* OR SUBMICRO*-ENGIN* OR SUBMICRO*-CHIP* OR SUBMICRO*-LAYER* OR SUBMICRO*-DIAMET* OR SUBMICRO*-ELECTRON*).

Query 2b: (NANOMET*-SCALE* OR NANOMET*SCALE* OR NANOMET*-LENGTH* OR NANOMET*-SIZE* OR NANOMET*-ENGIN* OR NANOMET*SIZE* OR NANOMET*-ORDER* OR NANOMET*-RANGE* OR NANOMET*-DIMENSION* OR NM-SCALE* OR NM-LENGTH* OR NM-SIZE* OR NM-ORDER* OR NM-RANGE* OR NM-DIMENSION*) NOT (WAVELENGTH* OR ABSORB* OR ABSORPT* OR ROUGHNESS).

Query 2c: (SUBMICRO*-SCALE* OR SUBMICRO*-LENGTH* OR SUBMICRO*-SIZE* OR SUBMICRO*-ORDER* OR SUBMICRO*-RANGE* OR SUBMICRO*-DIMENSION*) NOT (WAVELENGTH* OR ABSORB* OR ABSORPT* OR ROUGHNESS).

Articles excluded from Q2: (WAVELENGTH* OR ABSORB* OR ABSORPT* OR ROUGHNESS).

Query 3: (ATOM*-FORCE-MICROSCOP*) OR TUNNEL*-MICROSCOP* OR (SCANNING-PROBE-MICROSCOP*) OR (SCANNING-FORCE-MICROSCOP*) OR (CHEMICAL-FORCE-MICROSCOP*) OR (NEAR-FIELD-MICROSCOP*) OR (MOLECULAR-BEAM-EPITAXY) OR (MBE AND (EPITAX*

[11] For example, in another context, patent-publication lexical proximity, we used a weighting scheme favoring middle-low frequencies (log parabolic scheme) (Bassecoulard & Zitt, 2004).

OR GROW*)) OR QUANTUM-DOT* OR QUANTUM-DEVICE* OR QUANTUM-WIRE* OR COULOMB-BLOCKADE* OR COULOMB-STAIRCASE* OR LANGMUIR-BLODGETT.

Query 4: (SELF-ORGANI*D-GROWTH*) OR POSITION*-ASSEMBL* OR MOLECULAR-TEMPLAT* OR SUPRAMOLECULAR-CHEMISTRY OR DRUG*-CARRIER*.

Query 5: (DRUG*-DELIVER* OR DRUG*-TARGET* OR GENE-THERAPY OR GENE-DELIVER*) AND (POLYMER* OR PARTICLE* OR ENCAPSUL* OR CONJUGATE* OR SITE-SPECI* OR SITE-TARGET*).

Query 6: IMMOBILI* AND (DNA OR RNA OR MRNA OR RNAS OR TEMPLAT* OR PRIMER OR PRIMERS OR OLIGONUCLEOTIDE* OR POLYNUCLEOTIDE*).

Query 7: (POLYMER OR POLYMERS) AND (IMMOBILI* OR CO-IMMOBILI* OR COIMMOBILI* OR CONJUGATE* OR COMPOSITE*) AND (PROTEIN* OR ANTIBOD* OR ENZYME* OR DNA OR RNA OR MRNA OR RNAS OR POLYNUCLEOTIDE* OR VIRUS*).

Query 8: (MOLECUL*-SELF-ASSEMBL*) OR (SELF-ASSEMBL*-M*LAYER*) OR (SELF-ASSEMBL*-DOT*) OR ULTRAVIOLET-LITHOGRA* OR UV-LITHOGRA* OR PDMS-STAMP* OR SOFT-LITHOGRA* OR (SURFACE*-MODIF* AND (SELF-ASSEMBL* OR MOLECUL*-LAYER* OR ATOMIC*-LAYER* OR M*LAYER* OR MULTI-LAYER* OR MONO-LAYER* OR (LAYER-BY-LAYER))).

Query 9: (ENCAPSUL* AND VIRUS*) OR BIOMOLECULAR-TEMPLAT* OR MODIF*-VIRUS* OR VIRUS*-MODIF*.

Query 10: (PATTERN* OR SELF-ASSEMBL*) AND (ORGANI*ED-ASSEMBL* OR BIOCOMPATIB* OR BIO–COMPATIB* OR BLOODCOMPATIB* OR BLOOD-COMPATIB* OR CELLSEEDING OR CELL-SEEDING OR CELL-THERAPY OR TISSUE*-REPAIR* OR EXTRACELLULAR-MATRIX* OR EXTRACELLULAR-MATRIC* OR TISSUE*-ENGINEERING OR BIOSENSOR* OR IMMUNOSENSOR* OR CELL-ADHESION).

Query 11: SINGLE-MOLECUL* OR (SINGLE-ELECTRON*-TUNNEL*) OR MOLECUL*-MOTOR* OR MOLECUL*-BEACON* OR MOLECUL*-ENGIN* OR MOLECUL*-MANUFACT* OR BIOCHIP* OR DNA-CMOS OR FULLEREN*-TUB* OR FULLEREN*-PIP*.

Query 12 (*for exclusion*): PLANKTON* OR N*PLANKTON*OR M*PLANKTON* OR B*PLANKTON* OR P*PLANKTON* OR Z*PLANKTON* OR NANOFLAGEL* OR NANOALGA* OR NANOPROTIST* OR NANOFAUNA* OR NANO*ARYOTE* OR NANOHETEROTROPH* OR NANOPHTALM* OR NANOMELI* OR NANOPHYTO* OR NANOBACTERI*.

Query: (Q1–Q11) NOT Q12.

References

Bachmann, G. (1998). *Innovationsschub aus dem Nanokosmos. Technologieanalyse*. Düsseldorf: Bericht desVDI Technologiezentrum, Abteilung Zukünftige Technologie des Vereins Deutscher Ingenieure (VDI).

Bassecoulard, E., & Zitt, M. (2004). Patents and publications: the lexical connection. In W. Glaenzel, H. Moed, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research* (pp. 665–694). Dordrecht: Kluwer Academic Publishers.

Bonaccorsi, A. (2002, January 25–26). *Matching properties. Research regimes and institutional systems in science*. Paper presented at the science as an institution, the institution of science conference, Siena, Italy.

Bookstein, A. (1990a). Informetric distributions, part i: unified overview. *Journal of the American Society for Information Science, 41*(5), 368–375.

Bookstein, A. (1990b). Informetric distributions, part ii: resilience to ambiguity. *Journal of the American Society for Information Science, 41*(5), 376–386.

Braam, R. R., Moed, H. F., & van Raan, A. F. J. (1991). Mapping of science by combined cocitation and word analysis. I. Structural aspects. *Journal of the American Society of Information Science, 42*(4), 233–251.

Braun, T., Schubert, A., & Zsindely, S. (1997). Nanoscience and nanotechnology on the balance. *Scientometrics, 38*(2), 321–325.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and Isdn Systems, 30*(1–7), 107–117.

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: an introduction to co-word analysis. *Social Science Information, 22*(2), 191–235.

Darby, M. R., & Zucker, L. G. (2003). *Grilichesian breakthroughs: inventions of methods of inventing and firm entry in nanotechnology* (Working Paper No. 9825). Cambridge, MA 02138: NBER.

Egghe, L. (1990). The duality of informetric systems with applications to the empirical laws. *Journal of Information Science, 16*, 17–27.

Egghe, L. (2000). New informetric aspects of the Internet: some reflections – many problems. *Journal of Information Science, 26*(5), 329–335.

Egghe, L. (2005). The power of power laws and an interpretation of lotkaian informetric systems as self-similar fractals. *Journal of the American Society for Information Science and Technology, 56*(7), 669–675.

Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics. Quantitative methods in library, documentation and information science*. Amsterdam: Elsevier.

Fogelberg, H. (2003). The grand politics of technoscience: contextualizing nanotechnology. In H. Fogelberg & H. Glimell (Eds.), Bringing visibility to the invisible towards a social understanding of nanotechnology (pp. 29–47): Goeteborg University STS Research Reports 6.

Franks, A. (1987). Nanotechnology. *Journal of Physics E: Scientific instruments, 20*, 1442–1451.

Garfield, E. (1967). Primordial concepts, citation indexing and historio-bibliography. *Journal Library History, 2*, 235–249.

Glaenzel, W., Meyer, M., du Plessis, M., Thijs, B., Magerman, T., Schlemmer, B., et al. (2003). *Nanotechnology. Analysis of an emerging domain of scientific and technological endeavour (Intermediary results)*. B-3000 Leuven: Steunpunt O&O Statistieken, K.U Leuven.

Katz, S. J. (1999). The self-similar science system. *Research Policy, 28*(5), 501–517.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*, 10–25.

Kostoff, R. N., delRio, J. A., Humenik, J. A., Garcia, E. O., & Ramirez, A. M. (2001). Citation mining: integrating text mining and bibliometrics for research user profiling. *Journal of the American Society for Information Science and Technology, 52*(13), 1148–1156.

Laredo, P. (2002, November 6–9). *Six major challenges facing public intervention in higher education, science, technology and innovation*. Paper presented at the 4th triple helix conference, Copenhagen.

Leydesdorff, L. (2004). The university-industry knowledge relationship: analyzing patents and the science base of technologies. *Journal of the American Society for Information Science and Technology, 55*(11), 991–1001.

Malsch, I. (1997). *Nanotechnology in Europe: experts perceptions and scientific relations between sub-areas* (No. EUR 17710EN). Seville: European Commission – JRC, Institute for Prospective Technological Studies.

Marshakova, I. V. (1973). Document coupling system based on references taken from Science Citation Index. *Nauchno-TeknicheskayaInformatsiya, 2*(6.3) (in Russian).

Meyer, M. (2000). Does science push technology? Patents citing scientific literature. *Research Policy, 29*(3), 409–434.

Meyer, M., Persson, O., & Power, Y. (2001). *Nanotechnology expert group and eurotech data mapping excellence in nanotechnologies (No. Preparatory study)*. Brussels: EC, DG-Research.

Moed, H. F., & Vriens, M. (1989). Possible inaccuracies occurring in citation analysis. *Journal of Information Science, 15*, 95–107.

Murugesan, P., & Moravcsik, M. J. (1978). Variation of the nature of citation measures with journal and scientific specialties. *Journal of the American Society for Information Science, 29*, 141–155.

Naranan, S. (1970). Bradford's law of bibliography of science: an interpretation. *Nature, 227*, 631–632.

Noyons, E. C. M. (1999). *Bibliometric mapping as a science policy and research management tool*. Leiden: University DSWO Press.

Noyons, E. C., Buter, R. K., Hinze, S., van Raan, A. F. J., Schmoch, U., & Heinze, T., et al. (2003). *Mapping excellence in science and technology across Europe: nanoscience and nanotechnology* (Draft Report No. EC-PPN CT 2002-0001): EC.

NSF, Roco, M. C., & Bainbridge, W. S. (Eds.). (2002). *Converging technologies for improving human performance. Nanotechnology, biotechnology, information technology and cognitive science*. Arlington, Virginia: NSF/DOC-sponsored report.

Pao, M. L. (1993). Term and citation retrieval – a field-study. *Information Processing & Management, 29*(1), 95–112.

Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: theory, with application to the literature of physics. *Information Processing & Management, 12*, 297–312.

Rip, A. (1988). Mapping of science: possibilities and limitations. In A. F. J. v. Raam (Ed.), *Handbook of quantitative studies of science and technology* (pp. 253–273). Amsterdam: Elsevier.

Rousseau, R. (1990). Relations between continuous versions of bibliometric laws. *Journal of the American Society for Information Science, 41*, 197–203.

Rousseau, R. (1997). Sitations: an explanatory study. *Cybermetrics, 1*(1), 7.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New-York: McGraw-Hill.

Schmid, G., Decker, M., Ernst, H., Fuchs, H., Gruenwald, W., Grunwald, A., et al. (2003). Small dimensions and material properties. A definition of nanotechnology. *Europaeische Akademie Graue Reihe, 35*, 1–134.

Schummer, J. (2004). Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology. *Scientometrics, 59*(3), 425–465.

Small, H. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*(4), 265–269.

Small, H., & Griffith, B. C. (1974). The structure of scientific literatures i: identifying and graphing specialties. *Science Studies, 4*, 17–40.

Small, H., & Sweeney, E. (1985). Clustering the Science Citation Index using co-citations I – a comparison of methods. *Scientometrics, 7*(3–6), 391–409.

van Raan, A. F. J. (2000). On growth, ageing, and fractal differentiation of science. *Scientometrics, 47*(2), 347–362.

van Raan, A. F. J. (2001). Competition amongst scientists for publication status: toward a model of scientific publication and citation distributions. *Scientometrics, 51*(1), 347–357.

Zitt, M. (2005). Facing diversity of science: a challenge for bibliometric indicators – comments on a Van Raan's focus article. *Measurement: Interdisciplinary Research and Perspectives, 3*(1), 38–49.

Zitt, M., & Bassecoulard, E. (1996). Reassessment of co-citation methods for science indicators: effect of methods improving recall rates. *Scientometrics, 37*(2), 223–244.

Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2003). Bridging citation and reference distributions: part I. The referencing-structure function and its application to co-citation and co-item studies. *Scientometrics, 57*(1), 93–118.

Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: from cross-field to cross-scale effects of field-normalisation. *Scientometrics, 63*(2), 373–401.