Scientometrics (2015) 102:2223–2245
DOI 10.1007/s11192-014-1482-5

Meso-level retrieval: IR-bibliometrics interplay and hybrid citation-words methods in scientific fields delineation

Michel Zitt

Received: 23 October 2014 / Published online: 7 December 2014

Abstract In this position paper, we comment on various approaches to the delineation of scientific fields or domains, a typical prerequisite for a wide class of bibliometric studies. There is growing evidence that this meso-level, between micro targets of typical IR and large disciplines handled by macro-level bibliometric studies, takes full advantage of hybrid approaches. Firstly, delineation tasks gain to combine the a priori thinking of traditional IR, which typically involves clearly targeted expectations, and the a posteriori thinking of bibliometric mapping, where the decisions are built on external structuring of the domain in a wider context. The combination of the two ways of thought is far from new, with IR increasingly building on bibliometric networks for query expansion, and bibliometrics building on IR for evaluating and refining its outcomes. Secondly, delineation benefits from the multi-network perspective, which gives different representations of the scientific topics, usually all the more converging than the objects are dense and well separated. Focusing on two basic networks—words and citations—various sequences or combinations of operations are discussed. Bibliometrics and IR, especially when properly combined in multi-network approaches, provide an efficient toolbox for studies of domains delimitation. It should be recalled however that the context of such studies is often loaded with policy stakes that ask for cautious supervision and consultation processes.

Keywords Bibliometrics · Information retrieval · Science mapping · Field delineation · Hybrid textual-citation techniques · Query expansion

Introduction

Modern bibliometrics can be defined as the analysis of networks associated to scientific activity and knowledge exchanges, especially citations, linguistic relations (textual/

M. Zitt (✉)
Lereco U1134, SAE2 Department, INRA, Rue de la Géraudière, BP71627, 44316 Nantes Cedex 03, France
e-mail: zitt@nantes.inra.fr

semantic), partnerships of various kinds and institutional/territorial affiliations. The traditional model of bibliometric data deals with matrixes which depict the multiple correspondences between authors, documents and the structuring elements attached to these networks, such as cited items or words. It has been enlarged to new types of texts and media, and new types of relations often mimicking classical ones, for example hyperlinks. Applied bibliometrics offer quite a large spectrum: evaluation, positioning and mapping on a background of knowledge creation and diffusion models. The development of altmetrics promises a generalization of conventional tools and practices of scientometrics.

Information retrieval appears both more focused in its aim, providing users with the most relevant data at highest speed and lowest cost, and broader in scope: considering science-oriented IR only, it shares the playground (the science networks), the associated skew distributions (small worlds; fractal) and most of the mathematical and statistical toolkit with bibliometrics. It undergoes the same extension towards new diffusion media and networks.

Exchanges between the two areas, far from being cosmetic, have deeply shaped both of them. Four decades ago, Garfield (1967) citation indexing opened a new avenue to information retrieval and Kessler (1963) pioneer work on documents association matched a retrieval perspective and a mapping tool. The 1970s saw the development of clustering techniques in IR. In the late 1990s, we watched another example of the cross fertilization of bibliometrics and IR upon a new object, the web. Google and the forerunner Clever, taking advantage of the analogy of hyperlinks and citations, are for one part inspired from the seminal work of Pinski and Narin (1976) on "influence weights" in journals' networks. Influence weight, a powerful model of knowledge and information exchanges, was itself coined from a kind of iteration of Garfield's impact factors. The new algorithms of influence were applied back to journal networks and bibliometrics by Bergstrom (2007), following Palacios-Huerta and Volij (2004). Bibliometrics also require relevant datasets, privilege unified data (terms and concepts, actor's names) and permanently needs evaluation of relevance of outcomes, for example clusters and maps.

The bibliometric-aided retrieval and the IR applications of bibliometrics are rich and protean. We shall focus here on a natural crossroads of the two approaches, the meso-level. It encompasses the fuzzy range of scientific domains between narrow queries targets and academic disciplines: fields, areas of research, large topics.

A typical kind of bibliometric commissioned studies aims at positioning or evaluating actors in a domain judged strategic or emergent and promising. Finding acceptable bibliometric delimitation for the field and for its component topics also, is essential to the quality of the conclusions, since those sets are then used as baselines, for example for citation scores. Besides, scientists active in that domain expect a correct description of their own activity, hence the coverage requirements as to every research areas. Although bibliometric and IR paths may be hard to differentiate in this particular delineation task as in many others, we will here assimilate their association to the complementarity of an a priori process represented by IR, with explicit targets embodied in queries, and an a posteriori process privileged by bibliometric network approaches, for example delineation based on science maps and clustering. Section "Where a priori and a posteriori processes meet" will be devoted to the complementarity of those typical IR and bibliometric ways, section "Where citations and texts meet" to the complementarity of networks, especially texts and citations. It may be useful to recall that the positions reflected in this paper express a bibliometrician's point of view.

Where a priori and a posteriori processes meet

The meso-level context is usually more heavily loaded with political stakes than the narrow scope of micro-level retrieval or author-level bibliometric assessment. Take for example the "scientific fields" with academic background: these are hardly gathered under one univocal definition, and the task of delineation largely depends on the point of view:

epistemology: knowledge mix—shared paradigms and theories, methods, objects
sociology: visible or invisible communities of research sharing norms or interests
politics/policy: institutional framework with science actors, users and stakeholders
library science: relevant classification categories, journals, key concepts or key authors
bibliometrics: some dense and segregated areas in one or several science network(s)

Scientometrics and information retrieval may appear to convey some objectivism, trying to keep political interests and sociological controversies at arm's length, but they cannot escape the social determinants of the entities they describe in their theories and handle in their practice. Moreover, disciplinary effects are strong, influencing socio-cognitive networks and communities and organizations (Gläser 2010). Several classical problems in scientometrics, say citation field normalization, have to distinguish between "apples and oranges" cases—fields deeply different in their scientific practice—and "small apples and big apples" cases—where prerequisites of run-of-the-mill bibliometric studies apply and where appropriate normalization makes comparisons possible with limitations due to scale-dependence (Zitt et al. 2005). Literature on the case of socially sensitive areas and controversies (GMO, environmental issues) is abundant. Nomenclatures and associated delineation questions remain a touchy topic. The delimitation of academic fields or strategic domains with a strong institutional framework and funding bodies, is particularly sensitive to policy issues. Lastly, whatever the objectivism attached to quantitative analysis, final decisions are required, where the actors involved often redistribute cards.

Another aspect to consider is the complex nature of science networks. If scientific landscape were resembling Monument Valley, with its tabular relief and steep gradients, the delineation issue would be limited to some arbitrage, with such nicely separated mountains indeed easily deemed in or out the territory claim. Dense and relatively isolated themes do exist. In most cases however, the multidimensionality of cognitive and social relations, once watched through the dimension reduction power of data analysis, shows more complex relief and less steep gradients. Furthermore, the picture varies with the type of network used, and the type of setting applied, for example network weighting options.

This complexity and the overlapping character of scientific topics show that the question of "how to delineate a field" is not a trivial one that would already be solved by "ready-made IR" embodied in scientific nomenclatures. They are numerous, some institutional (coarse grain OECD and UNESCO) and others linked to databases (WoS, Scopus). The limitations of those general purpose frameworks, which must choose between short-term representativeness and reasonable stability of categories, soon appear when targeted delineation is needed. Nomenclatures are typically coarse grain—SCI/WoS was at a time exclusively journal-based—with hard classes not allowing for overlaps.[1] The most performing perhaps are the traditional classification schemes associated to such specialized databases as Medline, CAS or Econlit. Those classifications are often quite detailed and

[1] Amongst Thomson-Reuters nomenclatures, the "subject categories" of SCI classification allow for overlaps mostly in terms of journals or journals sections.

high performing, with assignments at the paper level, they can provide bricks for delineation tasks. For bibliometric purposes, most specialized databases suffer from being deprived of essential features of the Web of Science or Scopus, especially the citation index and the detailed affiliations of all authors.

Therefore, ready-made IR is generally insufficient or incomplete, and delineation processes involve specific IR querying tasks and/or bibliometric mapping.

IR approach

The archetypical approach of Information Retrieval is based on an “a priori” type of thinking: the mental model of expectations has to be structured beforehand, even fuzzily: the minimum requirement is that the mental model, whatever the degree of cognitive elaboration, can be translated into a set of queries, at least at the initial stage. The queries will undergo further learning and adjustments iterations. Without structured expectations, say in the form of a fuzzy cognitive representation of the field,[2] the queries will be limited to general formulas without any chance of good IR performance. The recall, especially, may be jeopardized by missing out key components of the field. In contrast, the typical bibliometric mapping conveys an “a posteriori” type of thinking. The structure and content of the database—or part of it—is mapped first and the outcome is confronted to the expectations in fine. For example, “self-assembly” or, nowadays, “graphene” will pop up as (set of) clusters, sparing the initial steps of the IR way. Breaking up with the deductive model, a posteriori thinking embodied in statistical induction has been held as the backbone of data analysis by one of its pioneers (Benzecri 1973).

The scale-up of IR queries is a traditional way to precisely delineate topics at all scales, including large fields. Virtually all information tokens from articles or other media have been mobilized for retrieval purposes or for studying a variety of relations between actors. The presence of assignments to nomenclatures or classifications, if any, is a first resource, with the limitations mentioned above. Tentative selection of Bradford core journals is a second one. Only moderate IR performance may be expected in complex domains and especially emerging ones, not being yet supported by a large editorial base. In association with journal selection or standing-alone, the standard delineation method adds conventional term queries from various bibliographic fields. Selection of core actors, sometimes with systematic protocols, is also classical. Coauthorship and especially citation data, from the bibliometric tradition, go along with conventional term-based IR. Other features of documents or authors, in the altmetric rationale, might be added in the retrieval toolbox.

Pure a priori thinking meets particular difficulties at the meso-level, with the exacerbation of the IR trade-off. Large scope queries (say the “nano” prefix if we wish to target nanosciences and technology) cannot be used alone, needing both restrictions and enhancements. It turns out that a collection of narrower queries (such as “self-assembly”, “quantum dots”, etc.) will be required for substituting or completing “nano” in order to encompass the scope of the domain. A priori thinking in delineation is typically associated with bottom–up assembly of partial queries with, in the case of a Boolean model, a wide use of the union operator. In other words, a preconceived breakdown of the domain is required. This step is usually carried out with the help of a panel of scientists/experts. The supervision of the process goes along with an evaluation of outcomes available for each sub-query and avoids black box effects. Compared with large scope queries (e.g. name of

[2] For a typology of IR models and the perspective of the “cognitive actor”, see Ingwersen and Järvelin (2005).

the field), better precision and recall potential is expected from the bottom–up path; compared with mapping by hard clustering, the overlaps of partial queries make the process locally robust.

There are also some downsides: experts' panels often show specialization biases, giving a poor collective a priori mental map of the domain, especially for multidisciplinary or controversial topics. Borders of domains create difficulties, hence threats on global recall. The multiplication of queries also needs time-consuming control of the frequent terms that are helpful for recall and harmful for precision. The most frequent ones are usually ruled out. The more varied the domain and the set of queries, the more ponderous and costly the supervision tasks and the adaptive processes. A global validation, established by sampling the whole literature retrieved, is another option.

Mapping for delineation purposes

The second family directly refers to the tradition of bibliometric mapping: bibliographic coupling (Kessler op.cit.), co-citation (Small 1973; Marshakova 1973) and journal–journal influence (Narin et al. 1976). Journal metrics rely on citations chains, citation exchange profiles (Leydesdorff and Cozzens 1993) and citation mutual exchanges (Bassecoulard and Zitt 1999) among other techniques. Many journal level classifications have been put forward over the past decades, the last ones with overlay facilities for positioning activities (Rafols et al. 2010). Other proposals use prior categories and expert judgments as seeds (Glänzel and Schubert 2003; Archambault et al. 2011), with reassignment of individual papers. Boyack and Klavans (2010, 2013) worked on maps and clustering with several granularities of the basic items (journals, papers) and of the aggregate categories. Their recent proposals add hybrid features (2010, 2013). Another recent paper-level classification focuses on symmetrized direct citations (Waltmann and van Eck 2012). High-quality delineation of fields cannot solely rely on journal-level classification, and this is still more conspicuous for emerging and complex domains.

In addition to query systems, two typical bibliometric exploitations of science networks address the delineation issues. The first one, the mapping of the science network on a scale—possibly all science—larger than the target domain, with a top–down zooming on the expected target areas, is self-standing. The second one consists in neighborhood exploration around a seed of retrieved literature identified beforehand.

Bibliometric mapping is an entry to field-delineation, with mental expectations of users/experts about the target domain projected a posteriori on a wider science landscape. The process is basically top–down [science→domain], with possibly a two-step process involving bottom–up reconstruction of the domain starting from a finer-grain top–down breakdown [science→subdomains→domain]. Mapping is usually associated with grouping procedures of some kind, if only to improve the readability of maps when zooming off. There are many types of rendering, from direct network representations optimized in some way to connected groups detected by data and graph analysis: factor/spectral analyses, classical clustering, fast algorithms of community detection such as Louvain's (Blondel et al. 2008), etc.

An advantage of a posteriori thinking is that the supervision task is reduced to the interactive comparison between the target and the landscape, less demanding in principle than the a priori conception of a query set. In terms of retrieval performance, an evident strength of a posteriori processes is that the risk of silence or noise on entire subdomains, at the scale(s) of observation picked, is drastically reduced. Clearly "out" subdomains will also be easy to detect (clustered noise), thus escaping threats on global precision. The discussion of the border of the domain especially, is made easier thus than with a conventional query system.

In addition, those displays that reflect the density gradients give an idea of the domain's local structure's effect on the robustness of delineation. If we go back to the low-dimensional geographical analogy, in/out decisions are easy to make for Monument Valley configurations (the "nano-objects" topic for example), in contrast with fuzzier structures. However, the rendering is methodology dependent. An obvious issue in textual maps, given the Zipfian distribution of words, is that for a given level of breakdown, frequent words should be shared amongst a number of clusters which tends to vary with the frequency. Hard clustering may force unique assignment with the help of various weighting systems, for example tf-idf.

The clustering options do matter for delineation when the conditions of the study involve groupings rather than scrutiny of the original network. Let us take the case of hard-clustering. Clustering from factor analyses can keep the overlaps but hard clustering remains widely used for reasons of computer efficiency. For instance, if the study is primarily based on words mapping, sources of overlaps are either ruled out beforehand (high frequency terms) or attached to a single cluster with possibly a low weight. Those issues are also met with citations, a bit less critical in that instance however, because of lower concentration. Hard-clustering forces the discrimination between sub-domains, and imposes clear-cut frontiers. Then, emphasis is more on precision than recall: a topic cannot extend to the "natural" border that an allowance for overlaps or multi-assignment would extend smoothly. The use of maps with hard-clustering for delineation purposes has ambivalent effects on recall: it guarantees that, in practice, no subdomain will be skipped (at the scale of observation), but might limit recall for individual sub-domains overlapping with neighbors or located on the frontier of the domain. Even for thematic clusters provided by native hard clustering, steps of enrichment can be added for expansion or overlaps handling. Early studies on clustering in IR already put into question the performances of top–down clustering.

Bridging a priori and a posteriori approaches

Clustering and data analyses are far from being the monopoly of bibliometrics. In IR, soon after Kessler's coupling, the "clustering hypothesis" (Jardine and van Rijsbergen 1971), stated that relevant documents tend to be more similar to each other than to non-relevant documents and tend to appear in the same clusters. This appeared as a radical alternative to the classical model, making it possible to test and combine top–down and bottom–up searches, taking advantage of the data structure and thus opening the way to mapping-based delineation The same principle was later tested on retrieved sets rather than static clustering on the collection. Traditional IR representations—Boolean, vector-space or probabilistic—can exploit terms interdependences and therefore the bibliometric networks.

A natural application is the query expansion processes. The archetypal method is perhaps "retrieval feedback", which identifies the terms beforehand, isolated or in interdependency (co-occurrences), specifically present in the most relevant documents retrieved. Evaluations are classically based on Rocchio (1971) and Salton and Buckley (1990); for a recent review and comparison of these models see Carpineto and Romano (2012). Various data analyses are efficient for extracting semantic concepts such as correspondence analysis and LSA (Deerwester et al. 1990) can also trigger further enrichment stages, if only by detection of synonyms. The processes used for analyzing the datasets can be applied to multi-networks, with examples in section "Where citations and texts meet".

Both IR and applied bibliometrics welcome mixed strategies with learning processes, adaptive queries and multistep protocols, with possible combination of supervised and automatic stages. Mapping, for example, may start from datasets of any origin. Usually based on papers datasets at some stage of elaboration, it can use queries datasets instead, for example by tracking term co-occurrences in user's queries (e.g. Ross and Wolfram 2000).

As far as field or topic delineation is concerned, typical IR a priori and bibliometric a posteriori strategies are seldom pure. The meso-scale particularly benefits from the converging toolboxes of the two approaches. A classical scheme of bibliometric study adds to a prior delineation task a cost-effective automatic breakdown into subdomains for actor's positioning, possibly refined by further cluster-based querying. In this case, the classification stage also helps to discard clustered noise. In cases where delineation is more complex, bibliometric mapping or network mobilization brings the efficiency of largely unsupervised processes at low cost, IR brings refinement where high precision and control are required. Neighborhood-based expansion, substitute or complementary to query expansion, typically completes the process. Multistep protocols are hardly avoidable, with chaining [IR step→network-based expansion] as well as [mapping step→IR-queries expansion]. Ingwersen (1996) works, especially on poly-representation, propose a theoretical foundation to the combination of cognitive points of views and illustrate the dialogue of IR tradition and bibliometric thinking, associating different document representations.

The transposition of those mechanisms to citations is straightforward, and there again the complementarity between traditional IR and network thinking has been recognized very early on. A retrieval enhancement may be sought by papers neighborhood in the network of documents, following Kessler's insights. The starting point may be the cited side as well, with detection of the cited cores. Protocols may rely on direct or indirect proximity (or n-path in a non-weighted network); require or not require clustering beforehand; and use or not use the same family of networks for the first stage and for the expansion or iteration(s). Larsen (2002) talked of a "boomerang" effect to discuss the construction of seeds and their citation expansion using a variety of ways. We will see examples in the next section.

Network-based expansion, as well as mapping, is not without limitations, especially the massive character of mapping and clustering, as opposed to finer and more intensive IR practices and semantic analyses. There are two traditional options for data analysis of bibliometric matrices, paper × items, paper × cited references or more generally structured items × structuring items, a dichotomy also found in IR applications. One is the direct classification of papers, through distances [paper × paper], in the rationale of bibliographic coupling—or lexical coupling; or else of direct citation linkages. The bibliographic coupling option exhibits certain advantages (see Glänzel and Czerwon 1996) but also shortcomings, since the underlying intellectual structure of the research hardly appears. The resulting relations between papers are holistic and the measured proximity index results from various semantic grounds.[3] An alternative path, that of [structuring

[3] Assume articles B and A share the theoretical background and C and A share the domain of application. In bibliographic coupling articles B and C may both be attracted by A on quite different semantic aspects, while without epistemic relation. The argument is already found in Martyn (1964). Even mitigated by statistical aggregation, it expresses the cost associated to the statistical efficiency of bibliometric clustering. The use of hard-clustering, simple and fast, worsens this limitation. An overlapping technique might classify A, once with B, once with C. IR scholars warned against the holistic character of several mapping techniques, source of noise including for query expansion purposes.

item × structuring item], either co-word or co-citation, gives a better although quite fragile connection with semantics (co-word), or with the intellectual base (co-citation), possibly suggesting primitive semantic interpretations. This approach is directly confronted with the skewness of word and cites distribution, with the challenge of overlaps mentioned above. On the other hand, primary clusters of structuring items (co-word themes, co-citation fronts), even produced by hard-clustering, can be transformed into secondary clusters of articles through multi-assignment of papers to themes or fronts, with facilities of interpretation.

Mapping and network-based expansion cannot totally skip supervision: any literature involved in bibliometric analysis has to be controlled at least at one stage, and often both at the launching stage and the final stage. Relevance indexes associated to the network metrics suggest solutions, which require final validation. For example, a sample of successive classes of documents ranked by relevance can be submitted to experts' evaluation for fixing a borderline. A few parallel processes (like words vs. citations, see next section) may bring a partial cross-validation independent from experts. Comparing bibliometric relevance and experts' opinions in a two-step hybrid delineation of genomics, Laurens et al. (2010) also suggested that in a hybrid sequential process, if the supervised word-based seed is held as the standard, the citation-based expansion might be pushed up to the marginal relevance of the seed, using for example the relevance of the last decile of the seed as an indicator.

More generally, we may consider the resources of data analysis, mapping techniques and topic modeling, a common substrate for IR and bibliometrics, as bridges for playing with a priori and a posteriori representations. On the one hand, modern IR makes a large use of clustering and mapping techniques to improve and expand queries and render their outcome. On the other hand, for each bibliometric set created at some stage by clustering or mapping operations, the equivalent IR steps that might have generated the set may (almost) be reconstituted by a kind of "retro-querying" process. Approximate reconstitution of data is inherent to factor analysis, for example correspondence analysis. Topic modeling through LSA, LDA and further refinements, search the probable model that might have generated the data (Papadimitriou et al. 1998). Less directly, automatic clustering including graph-based algorithms, with further specificity characterization of clusters, may be run in the same spirit. Variants of itemset mining uncovering association rules (Agrawal et al. 1993, with earlier forerunners), can handle NOT clauses (Cadot and Lelu 2011). All make the specific terms of each area explicit and may allow for the translation of the initial search into a new set of possibly more independent queries, with, usually, less trivial overlaps than at the first stage. The reconstitution of the initial system of queries is approximate, but at the expense of some noise, the translation makes some new associations and equivalences of terms apparent, and suggests paths for search expansion or restriction.

Figure 1 sketches some of the basic relations between the two paths, and the bridges available from the data analysis of literature retrieved by either approach. Besides, retro-querying has a similar function for associating views from different networks, exploiting the poly-representation of information retrieval or the bibliometric multi-network perspective.

To conclude this part, a pragmatic mix of IR and bibliometrics strategies is desirable for addressing field delineation. The difference between the favorite scale of IR, typically micro-level and therefore more idiosyncratic and prone to tailor-made solutions, and that of bibliometrics, the macro-level which is the preferential playground of mapping

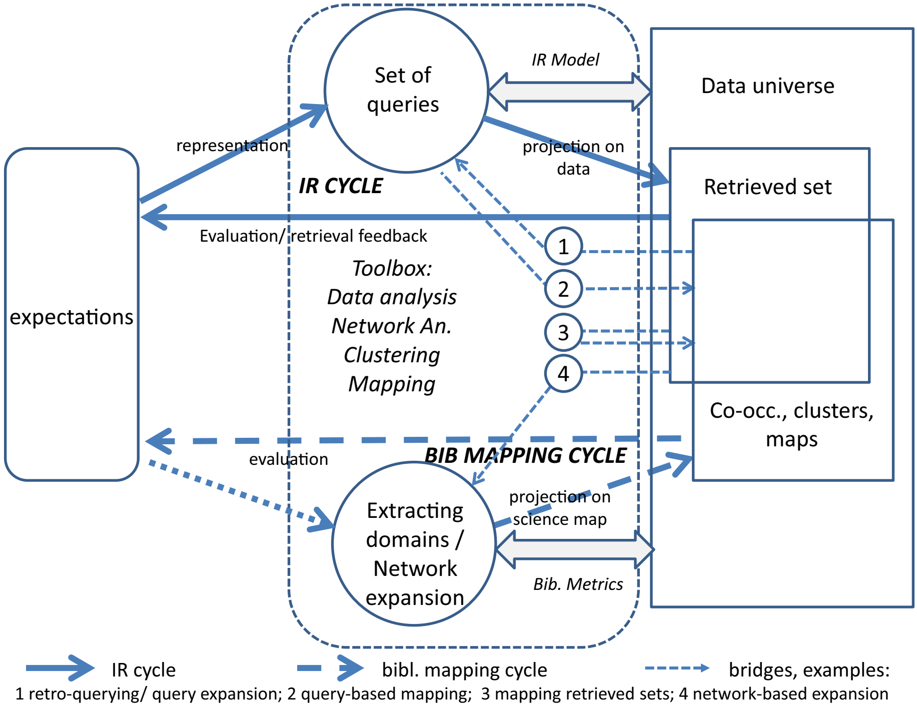


Fig. 1 Field delineation: Sketch of relations between IR and bibliometric approaches. The *mapping cycle* builds upon data to create and render networks and identify clusters. The clusters are confronted a posteriori with users' expectations (mental model: basic sketches or cognitive models) to reach a satisfactory delineation, which may need several runs. The process is either top–down, by extracting a domain's map from a larger representation, or bottom–up, by expansion of a seed. The *IR cycle* derives a set of queries from expectations. The more elaborate the mental model, the more detailed the queries. The retrieved set is evaluated on the basis of user's expectations, with several runs typically. The sets resulting from the delineation process (rectangles), retrieved by pure IR cycle and by pure mapping cycle do not totally coincide. The associated breakdowns also differ: the IR image is an overlapping collection of retrieved sets from a union of queries, while bibliometric data aims at ex-post structuring (clusters or spectral dimensions) of the universe. *Bridges* are provided by data analyses techniques in either direction. Starting from bibliometric delineation, retro-querying may help to refine expansion/restriction, possibly theme by theme (*1*). Symmetrically, the IR retrieved set can be mapped and clustered, for example for making latent dimensions or clusters explicit and questioning the initial queries arrangement (*2*–*3*). Both approaches allow for expansion or selection in multistep procedures, for example network-based expansion of an IR seed (*4*)

exercises,[4] is mitigated when they meet on the meso-scale of field-delineation. In addition, the fertilization between a priori (target-based) retrieval and a posteriori detection of areas in maps, or query expansion through proximity in the network, will benefit from a crossing with the multi-network approach.

Where citations and texts meet

The bibliometric delineation of fields and topics involves connections and representations reflected in various bibliometric networks. The idea that multiple entries provide

[4] we are indebted to an anonymous referee for stressing this point.

complementary views and that their local coincidence is a clue of relevance, is also quite common in IR, especially with the concept of poly-representation. Thematic studies in bibliometrics, also, most often start with such pragmatic combinations to reach an operational delineation of the domain under scrutiny: take the most cited authors, the main affiliations, the partnerships, the contents markers, etc. into query lists. Multi-network querying is therefore a run-of-the-mill operation in both information retrieval and applied bibliometrics. Particularly, an extremely wide class of bibliometric questions can be answered either by textual or by citation approaches (topic identification, characterization of emergence, static and dynamic mapping, diffusion processes, knowledge flows in science and more generally in the science-technology-innovation system), with different specifications and performances however.

An interesting use of "retro-querying", mentioned above, can take place at the early stage of field delimitation, in a general rationale of poly-representation. For example, let us start the study of a domain with a simple list of specialized journals. The content analysis through data analysis can translate this set into a ranked list of terms, or list of cited references, hence IR queries to enhance the set. There are various ways to take the best of each bibliometric network.

We shall focus here on some systematic attempts of combination for delineation purposes, limiting ourselves with the two networks generally considered good performers in indexing and mapping: document terms and citations. Obvious challengers are authors-based relations, especially authorship/co-affiliation, author co-citation and co-authorship. Authors' co-citation, successfully applied to fields description (White and Griffith 1981, Mc Cain 1983) is too coarse-grain to compete at the finer level—individuals often embrace a wide scope of topics, at a given time or successively—but presents distinct advantages from a sociological point of view by connecting to actor's network.

The relationship between citation and textual approaches has fed both pragmatic association and ideological confrontation. The introduction of citation indexing and citation mapping triggered the need for complementary textual information: in the heroic times, the Philadelphia teams tagged co-citation research fronts by "keyword +" (Garfield and Sher 1993), borrowed from cited articles; the influence weight (Pinski and Narin 1976) offered a new version of the citation network, which inspired many applications. At the same period, the capability of treating texts was soaring due to the development of computational linguistics on the one hand and the application of data analysis to texts, systematized by Salton's oeuvre and many authors such as Benzecri on the other hand (Benzecri et al. 1981 for the extensive application of correspondence analysis to vocabulary). Automatic query expansion, conceptualized in the 1970s, developed essentially from linguistic retrieval. In another research line, deBeaver and Rosen 1979) laid the first layers of collaboration studies, which are another gateway to social communities and IR (Mutschke et al. 2011). It became increasingly clear that mapping knowledge and its actors, or on a minor mode information exchanges, could rely on a few types of bibliometric networks associated to scientific activity, amenable to similar types of mathematical treatment based on local similarity or global metrics.

Competition also took place between the two points of view. The question arose of which network was the more appropriate for describing and understanding science, at a time (1980s) where citation evaluation, indexing and mapping were gaining interest. The British and French sociologists leading the "strong program in sociology of science" at the origin of the science studies movement and further developments such as the Actor Network Theory, defended the textual networks (Callon et al. 1983) against citations to represent knowledge on a background of actor's interests. Natural language appeared abler to depict science in action and especially in controversial areas, against "cold science" (Latour 1987) that citations might capture. The basic elements of network theory were soon enrolled for that purpose, while on the citation side de Solla Price, pioneer of citation networks (1965) had established the first network model equivalent to preferential attachment (1976), echoing Merton's Matthew effect. Patent-publication citation analysis and co-word approach also presented alternative paths for the description of science-technology relations (Narin and Noma 1985; Callon et al. 1991). Many variants of word-based scientometric maps were further developed, e.g. Noyons (1999).

Then computational linguistics and semantic networks also gave greater power to text analyses, still in competition with less sophisticated but massive statistic treatments, which benefited both from computer resources and new developments of spectral methods and fast clustering for large graphs, mentioned above. Citation techniques also underwent new developments, and efficient tools for mapping applications became increasingly popular (CiteSpace; Chen 2006). Research on normalization of citation networks has also been active and revival of influence measures for hyperlinks fed back citation analyses.

Lexical and citation characterization classically used in bibliometrics are appropriate for large statistical treatment. The basic resources at the paper level (bag of words, lists of references) prove rich in nuggets when transformed by LSA or mapping methods. However, those exploitations look shallow when compared to sophisticated linguistic methods or (heavy) application of semantic networks. A particular example of a bridge between semantics, science models and bibliometrics is the research line of citation in context.

The "citations in context" track was initiated in co-citation studies (e.g. Small 1980) which are a natural space to connect referencing, intellectual base and linguistic aspects. The linguistic and semantic analysis of citations contexts has various objectives: contribution to the classical issue of citations types or motives (Teufel et al. 2006), cross analysis of the contents of the citing or the cited documents, classification of cited or citing documents after the citation context (Ritchie et al. 2008), fine-grain relation of citation contexts and abstracts terms (Liu and Chen 2013), exploration of new dimensions of scientific texts (Small 2011). Some of these advances influence citation techniques in return. An example is the improvement of co-citation accuracy (Elkiss et al. 2008, Callahan et al. 2010). Citation in context encompasses a wide scope of ambitions, from simple context visualization on citation engines to insights into dynamics of science. Citation in context studies are amongst the gates to access the mental and social processes of research in action, associating language and communities' life.

In parallel, new impulsions for combining various networks appeared. On the practical side, the culture of data-mining encouraged mixes between several networks for pragmatic purposes (Kostoff et al. 2001). On the theoretical side, the developments of network theory, especially the Watts–Strogatz small world and the Simon–Price–Barabasi attachment models (Barabasi and Albert 1999), found a privileged application in large-scale scientific networks, especially co-authorship and citations. A variety of models in relation with emerging structures that shape the landscapes of science was proposed in H. Simon's wake (Börner et al. 2003, Rosvall and Bergstrom 2008, Carayol and Roux 2009, Eom and Fortunato 2011, Watts and Gilbert 2011). Most privilege the author and/or the article entry but the interplay of co-authorship, citation and linguistic networks is increasingly gaining attention: relations between contents and actors' positions (Gilbert 1997; Roth and Cointet 2010), between citations and co-authorship, and any or both of these with texts (see for example Mutschke and Quan-Haase 2001). The detour by those various dimensions of science networks dynamics (Börner et al. 2011, Scharnhorst et al. 2012) might be a step, awaiting for more powerful models of authors and community behavior able to unify these diverse representations.

Word and citation networks: convergence and divergence

If we go back to applied bibliometrics and especially delineation, the question arises of convergences or divergences of lexical and citation approaches.

Both the textual contents and the cited reference are the results of authors' choice in their community context. Both involve a mix of scientific and social aspects: words and cited references are community markers and reflect the sociability of invisible colleges. A large body of literature (see Bornmann and Daniel's review 2006) has been devoted to the citation behavior, including Cronin's classic (1984). Whatever their determinants can be, Merton's rewards, Small's symbolic beacons, Gilbert's persuasion tools (1977) or Latourian interests—quest for immunity, enrolment of colleagues, return of favor—the references mainly point towards the thematic groups where partners and gatekeepers are found. On the text side, rhetoric and jargon expressing community habits, general words voicing interests, rejoin focused technical terms to define topics contextually. A substantial amount of convergence between texts and citations is therefore expected.

However, the interplay of linguistic and referencing aspects is complex and varied. The particularism of communities may focus on the intellectual repertoire (preference for citations to the same school of thought, to the same national corpus, etc.) while the vocabulary remains universal; or on linguistic factors only, due to different traditions in terminology, while the intellectual base is largely shared. In such cases deep divergences between text-based and citation-based mapping may locally appear. Other well-known differences are potential sources of divergence.

By nature, citations are diachronic, words primarily a-chronic. The same type of formalism applies (matrices docs/items and derived, family of skew distributions, network modeling). Basic matrix transformations lead to word associations/lexical coupling and citation equivalents. The opposition between a-chronic words and dated cited articles is mitigated in symmetrical forms (research fronts citing co-cited cores, bibliographic coupling) so that for example, dynamic chaining of research fronts on the one side (already a concern at the heroic times of ISI's Atlas of Science), of word-based clusters on the other (Chavalarias and Cointet 2013) are akin. The citation way adds explicit information of time-lag between cited and citing sets, and related immediacy/obsolescence analysis. Beyond classical dating of words, co-word, or clusters in time series, natural language analysis opens toward subtler analyses of word transformations in a scientific context (Polanco et al. 1995).

Statistical properties are different: beyond their common background of hyperbolic distributions, word distributions (Zipf-Mandelbrot) are more concentrated and less «complex» than their citation counterpart. The parameters of citation distributions are modulated by the citation-windows. As to co-occurrences, word-relations matrices are less sparse and likely to be noisier than citations relations.

The difficulties of text approaches are also the downside of the richness and versatility of natural language—even when restricted to a unique case, say international English. Polysemy, metonymy, synonymy, stylistic devices and disciplinary jargon are well-known traps of linguistic difficulties that the users—and the bibliometricians—have to cope with. Those language features interfere with the general translation issues between information systems language and user's formulation. Blair (2003) highlighted two aspects of this retrieval problem: failure of description of categories or documents that might be described by virtually infinite indexing—a threat on recall—and failure of discrimination that threatens precision or saturates the user. Many authors stress the cost of reaching unification of terms or concepts, including with general data analyses like CA, LSA, LDA or targeted disambiguation/unification tools. Early results already showed the performance of citation indexing in comparative retrieval tests, and especially the interest of cross retrieval (McCain 1989; Pao 1993). However, textual information preserves its advantages of availability, intuitiveness, and easy handling in simple or complex search.

Normalization of distances using various metrics also transforms valued networks with thresholds. The Zipfian distribution of terms is particularly challenging, creating massive cluster overlaps ("pyramidal" classes). With little normalization or neutralization of high-frequencies, lexical groupings tend to be driven by generic terms. The effect of various weighting options in function of node degree (inclusion index, equivalence index and Ochiai, probability index) on the configuration of co-word networks with threshold was already stressed in the 1980s.

Citation relations are less evocative than a good lexical map, but adding a layer of verbalization to the titles of a co-cited core or a bibliographic coupling cluster yields a similar and often more precise description. Another great capability of citation is the historical chaining of direct citation, with the permanent research track on citation clusters time-chaining since the ISI' Atlas of Science (e.g. Chen et al. 2010) and the revival of citation historiography (Garfield et al. 2003). A practical advantage, as long as a realistic level of imprecision is accepted—with respect to matching errors anyway—citation relations can be largely automatized, thus limiting costs of supervision. As a result of multi-network or poly-representation hypotheses, some issues typical of one representation can received a solution from the other. For example, citation-based clustering can solve a typical problem of natural language (synonymy) insofar as lexical and citation networks approximately coincide upon particular topics. Polysemy in a wide sense is a trickier issue, associated to several configurations: homonymy, generic highly connected terms, bridge-concepts between a few topics. Citation-based classifications are less sensitive to homonymy if the basic matching of references is correct, and a local comparison of clusters may detect a word-based clustering failure due to polysemy.

However, citations are not spared shortcomings. As a rule, biases of citations are less harmful in mapping applications than in citation scoring (impact, composite indexes). In mapping, relevance towards topic areas, mentioned above, rather than towards individual contributions, matters.[5] Other down-sides are more severe. The bandwagon effect in citation behavior tends to create spurious cliques in native co-citation networks, possibly hindering the discriminating power of citation relations. The inflation of the number of references in authors' practice, which is a long-term trend (Larivière et al. 2008), also brings noise to conventional citation clustering. The disciplinary insertion affects the number of references ("propensity to cite"), to such an extent that correcting for the length of references lists is an efficient form of partial field-normalization (Zitt and Small 2008) and knowledge-flows weighting, while the same principles of correction plus recursion of citation linkages define "influence measures" (Pinski and Narin 1976).

Application: combining and comparing words and citations in field mapping

To a certain extent, changing the type and parameters of the network is like observing the universe in various wavelengths. Perhaps, a fragmented landscape like Monument Valley

[5] Even Latourian citations or negative citations do not add much noise to co-citation topics.

will remain visible on a wide spectrum—the various types and settings of analysis, while smoother structures evoking the landscape of the Kent will be less robust towards the type of analysis. For words and citations, a reasonable conjecture is that a given object both dense and segregated tends to be retrieved by both ways. For example, in the "nano" domain, if we choose a radical reduction of dimensionality, e.g. a hills and valleys rendering, the "nano-objects" area will evoke table mountain with steep borders, whereas the area of theoretical substrate of nano research will exhibit more complex arrangements and low gradients. Now if we take a cross-representation lexical × bibliographic coupling of clusters, the delimitation of "nano-objects" for example, will tend to superpose as a table mountain; less dense and highly connected areas are more sensitive to the type of network picked and to the particular metrics and clustering technique employed. Generally speaking, a set of clustered papers belonging to a strong overlap of a word-based cluster and a citation-based may be considered as a strong form, in a rationale already present in the first comparisons of McCain (1989) on term vs. citation indexing. The power of poly-representation to induce relevance from convergence appears modulated by the local density of the networks. In less dense areas that sociology of translation associated to emerging or declining phases, groupings are technically more sensitive to methodology, so that the linkage between convergence of representations and relevance of clusters may be obscured.

Delineation through hybrid word-citations methods: various strategies

Many sequences and combinations have been explored by scholars. Millions of Google users benefit from hybrid IR processes every day. In spite of a large literature devoted to the PageRank algorithm itself starting with Brin and Page (1998), the detailed combination of multi-network operations is not documented, so that the more advanced techniques are not necessarily on the table. Only some quite basic multistep and multi-network combinations, focused on mapping and delineation, are envisioned here:

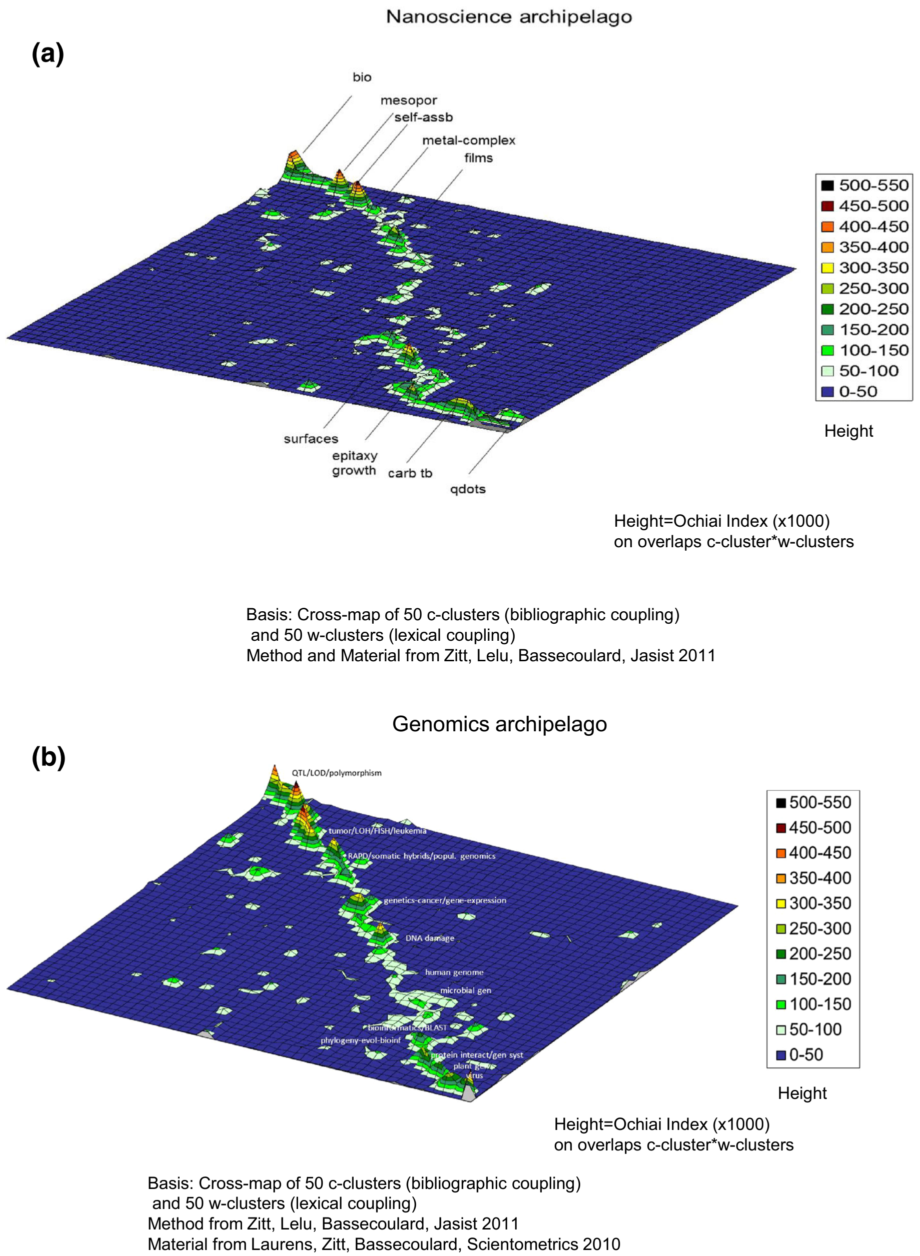
The first type of combination runs the two processes, texts and citations, in parallel as long as possible, allowing for distinct sociological interpretations. At the opposite end, the full hybrid treats terms and citations as miscible information tokens from the start. The former appears more sociology-compatible, the latter conveys a strong informetric or data-mining posture. In the middle stand the sequential hybrid processes, starting with a textual stage and proceeding with a citation stage, or the reverse (Table 1).

Parallel hybrid

It consists in separate exploration of the lexical content and the citation content of the same corpus, using appropriate metrics and similar clustering method, so that the final outcomes can be compared and/or combined. A typical application bears on comparison of breakdowns from a corpus already known, submitted to a parallel clustering operation. For example, in (Zitt et al. 2011; Laurens et al. op.cit.), using bibliographic coupling and lexical coupling with a particular document × document metrics, we obtained two partitions—one word-based, the other citation-based. Then we built the cross-clusters matrix and finally reordered it on its first dimension. The outcome is a matrix where heaviest intersections of c-clusters and w-clusters form a roughly diagonal zone, which can be directly transformed in an archipelago pseudo-map (Fig. 2).

Table 1 A few hybrid delineation processes: terms-citations

	Parallel	Full hybrid	Sequence cit-terms	Sequence terms-cit
Type	Compare/combines two self-standing and consistent representations	Informetric posture: mixed metrics from the beginning		
Pros	"Sociologically correct"; capability to compare; cross-validation	Technical flexibility	Can enrich citation clustering	Logical sequence of supervised process on words and automatic citation expansion
Cons	Heavy; sensitivity to differences in distribution, may benefit from correction	Mixes of two quite different phenomena	Difficulty to start from a citation-based seed without a prior exploration step; tuning of expansion	Fine tuning of expansion



Basis: Cross-map of 50 c-clusters (bibliographic coupling)
and 50 w-clusters (lexical coupling)
Method and Material from Zitt, Lelu, Bassecoulard, Jasist 2011

Basis: Cross-map of 50 c-clusters (bibliographic coupling)
and 50 w-clusters (lexical coupling)
Method from Zitt, Lelu, Bassecoulard, Jasist 2011
Material from Laurens, Zitt, Bassecoulard, Scientometrics 2010

Even slightly underestimated by the imperfect optimization of the reordering and the fact that whole papers, and not narrow citation contexts, are considered, the convergence of the c-partition and the w-partitions is pretty strong in the two cases shown, nano-sciences and genomics, but not to the point that a single one is sufficient:

◀**Fig. 2** Archipelagoes of science unveiled by cross-maps. A double 50-clusters breakdown was obtained by a variant of k-means, Axial k-Means (AKM, Lelu 1994), yielding two partitions of the set of papers, amenable to a cross-table of clusters intersections between c-clusters (bibliographic coupling rationale) and w-clusters (lexical coupling). The intersection matrix was then reordered by block-modeling to highlight the correspondences between the two kinds of clusters, on a roughly diagonal area—for an early application of block-modeling to bibliometric clusters, namely cocitation clusters and their social structure, see Mullins et al. (1977). Sinuous deviations (snake shape) to the diagonal are due to cluster-size distribution. The display makes apparent the relative strength of intersections (z axis). Figure 2a derives from the original map on nanosciences (2 dimensions) found in Zitt et al. 2011; a similar figure is shown for the genomics field (Fig. 2b). Differences between the c and w partitions are of several kinds: non optimal reordering missing some groups, separations which vanish with a change of scale, unreducible divergences due to contrasting informetric or sociological properties

– This gives evidence of the overall robustness of bibliometric clustering towards the type of network, lexical or citation. The narrower the diagonal zone, the more convergent the two approaches.
– Groupings at various scales of the strong forms (heavy intersections in relative terms) are suggested.

Direct application to delineation of sub-areas can exploit the granularity of intersections rather than that of clusters.[6] This may be extrapolated to the delineation of large fields, starting from larger oversets (all science in a limit case), by comparing partitions obtained for example by bibliographic and lexical coupling.

FULL HYBRID: combined metrics

The structuring/clustering of fields using a common metrics mixing citation and term distances at the finer grain level, from the start, is also a promising path (van den Besselaar and Heimeriks 2006). Statistical differences between word and citation distributions can be reduced through a proper normalization of the similarity measures (Janssens et al. 2008; Liu et al. 2010). The evaluation of hybrid metrics has just begun (Ahlgren and Colliander 2009). Boyack and Klavans (2010 op.cit.), on a large dataset, observed that even a “hybrid naïve” coupling outperformed pure bibliographic coupling.

Purely informatic methods by-passing words and citations and directly treating codes (character n-grams on text flow, compression) for calculating generalized text distances, are ultimate stages of hybridization, dissolving both terms and reference chains in signals. Such black boxes are deprived of semantic or bibliometric interpretation, but can reveal quite efficient.

Sequential hybrids

Limiting ourselves to seed-expansion series:

(a) CITATION → TERMS

We mentioned above the tradition of completing citation objects by textual tagging. The question of the validity of co-citation research fronts for research evaluation, due to

[6] High-precision is expected from a strategy focused on strong forms—heavy intersections—with check of the specific words/references of the native clusters c and w; high-recall is expected from a strategy based on the full content of c and w clusters with strong overlaps. Intermediary strategies can focus on intuitive groupings of areas along the diagonal sequence.

their poor coverage of literature, triggered further developments on retrieval and the means to foster it, possibly with the help of texts. For example, the enhancement of co-citation coverage by two-step expansion could be controlled by lexical means (Zitt and Bassecoulard 1996). Braam et al. (1991) developed a systematic complementation of co-citation clusters coverage by lexical means, a first operational example of hybrid delineation. The citation → text analysis sequence keeps being explored for other purposes. Recently Boyack and Klavans (2013), while preserving a co-citation breakdown of science, introduced textual metrics for large-scale displays of inter-clusters relations.

(b) TERMS → CITATIONS

The perspective is reversed. The remote ancestor is a classical application of citation indexing, when title words or keywords+ were used to query a citation index for collecting papers on a given (set of) topics. We focus here on a two-step protocol where the first step is based on IR queries on word relations, and the second step based on a network exploration on citation relations. There is some analogy with the “boomerang effect” quoted above. We shall take the example of the lex+cite process explored in Zitt and Bassecoulard (2006) especially for emerging or transverse domains, where classical methods tend to fall short.

The rationale is simple: words and citations are complementary; starting a multistep process with experts’ help is easier with word queries; unsupervised procedures are safer on citations, with proper precautions, than on words. A two-step protocol was set up

- on a supervised precision-oriented phase: standard queries: terms, journals → seed
- and a non-supervised recall-oriented phase: enhancement by citations → sister literature sharing the same specific intellectual base.

Term-based query expansion is hardly manageable without supervision or heavy contribution from external semantic resources. However, the progress in IR techniques erodes Salton’s reservations about the risks of query expansion. We argued that automatic enrichment by citations was safer to use in order to exploit the full potential of the complementarity between a priori and a posteriori processes. In practice, the lex + cite method proved efficient under general requirements for citation analysis, especially not too scarce reference lists. It can be global or rely on cluster-by-cluster enrichment if a previous breakdown is available. Depending on the type and settings of expansion, those techniques can both improve the recall on existing subdomains and rescue missing ones. Besides the recall-oriented aim, these protocols may also enhance precision through the detection of noisy subdomains or marginal literature.[7]

Naturally, most elaborate and iterated sequences may be imagined to take full advantage of hybridization. Hybrid approaches meet the same issues than those exploiting a single network: in absence of immanent deem on which document is relevant, the terms of reference of recall and precision are a matter of negotiation and incertitude. Protocols of experts’ guidance for evaluation purposes are required. Cross-validation of parallel processes, and even in some cases of sequential processes (Laurens et al. 2010) may alleviate the burden of multistep external validation.

[7] for example by ruling out papers without a given number or proportion of specific references.

Discussion: conclusion

Topic delimitation at micro-level appears as a recurring issue in applied bibliometrics or webometrics. Field or meso-scale delineation is not just a scale-up version of topic delineation. High-performance retrieval is required especially when further detailed scrutiny of subareas is planned, with questions of balanced representativeness of these subareas. It also involves data analysis methods along traditions of mapping and clustering, revived through fast algorithms of community detection and topic modeling.

The idea of bibliometric delineation does not assume any essentialism: the definition and delimitation of scientific fields encompass several dimensions, and rules of the game should be stated. Science is massively interconnected, whatever the network, so that again some rule of specific relevance or relevance/cost should be made explicit to reach reasonable borders. Maps warn against splitting dense objects or joining remote entities and suggest groupings and borders, with sensitivity of results to the metrics and aggregation methods. IR can alert on recall-precision performances once the criteria of reference is stated. However, quantitative analysis cannot dictate delineation solutions.

Let us summarize:

– Delineation of fields is a challenge in applied bibliometrics, in many cases where fast techniques such as core journals selection fall short.
– Progress in IR on the one hand, of mapping and clustering/topic-modeling techniques on the other hand increase the technical resources for delineation tasks.
– Combinations of “a priori” and “a posteriori” processes along multistep protocols can take the best of each: elaborate queries with large experts involvement, and strength of unsupervised mapping.
– A classical IR heritage, the retrieval feedback, gives a pivotal role to the analysis of the retrieved set in further steps. The rationale may be generalized to the shift of granularity or network along delineation processes (journal/terms, terms/citation); and to supplemental IR query steps of prior outcomes of mapping.
– Combinations of linguistic, citation and other relations in a multi-network perspective, rejoining IR poly-representation concept, harness the properties of each network at best. From the differences previously suggested between words and citations, we guess that the textual pathway is more intuitive but more demanding in expert supervision or semantic resources. Citations escape the ambiguities of natural language and can stand automatic treatments more easily. New investigations are needed to compare the efficiency of text, citations and authors’ communities in mapping and network-based delineation.
– The conjecture that dense and segregated areas tend to superpose whatever the type of network, especially words and citation, while local divergences are more likely to appear in low gradient areas, may alleviate delineation tasks.
– There are many ways to associate those cultures in hybrid delineation protocols: parallel process, full hybridization, sequences of textual and citation analysis of various types. Those combinations often convey a disciplinary point of view: full hybrid metrics reflect a strongly informetric or data-mining posture, with a reduction to information tokens. Parallel processes, more “sociologically-correct”, allow for socio-cognitive interpretation, and so for series protocols.
– As to sequential processes, we pleaded for supervised lexical approach first, completed with citation expansion with little supervision, an efficient and relatively simple arrangement.

– Citations in context narrow the bridge between citation and semantics, and open towards a better understanding of science in action. Further developments are expected from multi-networks of science, a detour through diversity of representations awaiting for strong unified representations.

Eventually, understanding the landscape of science in its multi-network aspects, an endeavor promised by research on dynamics of science, will give enhanced theoretical support to the current practical toolbox of field delineation. The discussion of in-topics and out-topics on a bibliometric map, or in parallel the standards behind the F-scores recall that decision-aided techniques are confronted to disciplinary traditions and institutional positions with non-explicit criteria. The methods of supervision and consultation suitable for revealing preferences, limiting specialization biases and handling policy stakes are as important in the practice of field-level studies. In a famous Daoist allegory, a butcher keeps the same knife for years without need for sharpening it, because he constantly follows the least-effort lines. The ideal cut-off for scientific domains and fields is certainly more challenging, even with Daoist equipment, when one comes to "bibliometrics in social context".

Acknowledgments The author thanks Alain Lelu, Université de Franche-Comté and Loria, Nancy, Elise Bassecoulard, formerly Inra-Lereco, and anonymous referees, for helpful remarks; Patricia Laurens and Antoine Schoen, ESIEE, Marne la Vallée, for permission to use the genomics map, from our previous co-work.

References

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD,* 207.

Ahlgren, P., & Colliander, C. (2009). Document-document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics, 3*(1), 49–63.

Archambault E., Beauchesne O. H., & Caruso J. (2011) Towards a multilingual, comprehensive and open scientific journal ontology, in *Proceedings* 13th ISSI *Conference*, Durban, South Africa.

Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509–512.

Bassecoulard, E., & Zitt, M. (1999). Indicators in a research institute: A multi-level classification of scientific journals. *Scientometrics, 44*(3), 23–345.

Benzecri, J. P. (1973) La place de l'a priori, *Encyclopedia Universalis*, *17*, Organum, 11–24.

Benzecri, J. P., et al. (1981). *Pratique de l'analyse des données : Linguistique et lexicologie*. Paris: Dunod.

Bergstrom, C. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *College & Research Libraries News, 68*(5). www.ala.org/ala/acrl/acrlpubs/crlnews/backissues2007/may2007/eigenfactor.cfm.

Blair, D. C. (2003). Information retrieval and the philosophy of language. *Annual Review of Information Science and Technology, 37*, 3–50.

Blondel V. D., Guillaume J. L., Lambiotte R., & Lefebvre E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), 10008.

Börner, K., Chen, C. M., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology, 37*, 179–255.

Börner, K., Glänzel, W., Scharnhorst, A., & van den Besselaar, P. (2011). Modeling science: studying the structure and dynamics of science. *Scientometrics, 89*, 347–348.

Bornmann, L., & Daniels, H. D. (2008). What do citation counts measure? A review of studies on citation behavior. *Journal of Documentation*, *64*(1), 45–80.

Boyack, K. W., Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *JASIST*, *61*(12), 2389–2404.

Boyack, K., & Klavans, R. (2013). Creation of a highly detailed, dynamic, global model and map of science, forthcoming in *JASIST*. doi:10.1002/asi.22990.

Boyack, K., Small, H., & Klavans, R. (2013). Improving the accuracy of co-citation clustering using full text. *JASIST, 64*(9), 1759–1767.

Braam, R. R., Moed, H. F., & Van Raan, A. F. J. (1991). Mapping of science by combined co-citation and word analysis. I Structural aspects. *JASIS, 42*(4), 233–251.

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.

Cadot M., & Lelu, A. (2011). Combining Explicitness and Classifying Performance via MIDOVA Lossless Representation for Qualitative Datasets. *International Journal on Advances in Software*, *5*(1–2), 1–16.

Callahan, A., Hockema, S., & Eysenbach, G. (2010). Contextual co-citation: Augmenting co-citation analysis and its applications. *JASIST, 61*(6), 1130–1143.

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.

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics, 22*(1), 155–205.

Carayol, N., & Roux, P. (2009). Knowledge flows and the geography of networks: A strategic model of small world formation. *Journal of Economic Behavior & Organization, 71*(2), 414–427.

Carpineto, G., & Romano, C. (2012). A survey of automatic query expansion in information retrieval. *ACM-CSUR, 44*(1), 1.

Chavalarias, D., & Cointet, J. P. (2013). Phylomemetic patterns in science evolution—The rise and fall of scientific fields. *PLoS ONE, 8*(2), e54847.

Chen, C. M. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *JASIS, 57*(3), 359–377.

Chen, C. M., Ibekwe-Sanjuan, F., & Hou, J. (2010). The structure and dynamics of co-citation clusters: A multiple-perspective co-citation analysis. *JASIST, 61*(7), 1386–1409.

Cronin, B. (1984). *The citation process; The role and significance of citations in scientific communication* (p. 103). London: Taylor Graham.

de Beaver, D., & Rosen, R. (1979). Studies in scientific collaboration. Part II. Scientific co-authorship, resrach productivity and visibility in the French Scientific Elite, 1799–1830. *Scientometrics, 1*(2), 133–149.

Deerwester, S., Dumai, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *JASIST, 41*(6), 391–407.

Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., & Radev, D. (2008). Blind men and elephants: What do citation summaries tell us about a research article? *JASIST, 59*(1), 51–62.

Eom, Y. H., & Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PLoS ONE, 6*(9), e24926. doi:10.1371/journal.pone.0024926.

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

Garfield, E., & Sher, I. H. (1993). Keywords-Plus(Tm) -Algorithmic derivative indexing. *JASIST, 44*(5), 298–299.

Garfield, E., Pudovkin, A. I., & Istomin, V. S. (2003). Why do we need algorithmic historiography? *JASIST, 54*(5), 400–412.

Gilbert, G. N. (1977). Referencing as persuasion. *Studies of Science*, *7*, 113–122.

Gilbert, N. (1997). A simulation of the structure of academic science. *Sociological Research Online, 2*(2), 3. http://www.socresonline.org.uk/socresonline/2/2/3.html.

Glänzel, W., & Czerwon, H. J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics, 37*(2), 195–221.

Glänzel, W., & Schubert, A. (2003). A new classification of the science fields and subfields designed for scientometric evaluation purposes. *Scientometrics, 56*(3), 357–367.

Gläser, J., Lange, S., Laudel, G., & Schimank, U. (2010). The Limits of Universality: How field-specific epistemic conditions affect authority relations and their consequences. In R. Whitley, J. Gläser, & L. Engwall (Eds.), *Reconfiguring knowledge production: Changing authority relationships in the sciences and their consequences for intellectual innovation* (pp. 291–324). Oxford: Oxford University Press.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation, 57*(6), 715–740.

Ingwersen, P., & Järvelin, K. (2005). *The turn: Integration of inversion seeking and retrieval in context* (p. 436). Berlin: Springer.

Janssens, F., Glanzel, W., & De Moor, B. (2008). A hybrid mapping of information science. *Scientometrics, 75*(3), 607–631.

Jardine, N., & van Rijsbergen, C. J. (1971). The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval, 7*, 217–240.

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. *JASIST, 52*(13), 1148–1156.

Larivière, V., Archambault, E., & Gingras, Y. (2008). Long-term variations in the aging of scientific literature: from exponential growth to steady-state science (1900–2004). *JASIST*, *59*(2), 288–296.

Larsen, B. (2002). Exploiting citation overlaps for information retrieval: Generating a boomerang effect from the network of scientific papers. *Scientometrics, 54*(2), 155–178.

Latour, B. (1987). *Science in action: How to follow Scientists and Engineers through society*. Cambridge: Harvard University Press.

Laurens, P., Zitt, M., & Bassecoulard, E. (2010). Delineation of the genomics field by hybrid citation-lexical methods: Interaction with experts and validation process. *Scientometrics, 82*(3), 647–662.

Lelu, A. (1994). Clusters and factors: Neural algorithms for a novel representation of huge and highly multidimensional data sets. In E. Diday & Y. Lechevallier (Eds.), *New approaches in classification and data analysis* (pp. 241–248). Berlin: Springer.

Leydesdorff, L., & Cozzens, S. E. (1993). The delineation of specialties in terms of journals using the dynamic journal set of the science citation Index. *Scientometrics, 26*, 133–154.

Liu, S., & Chen, C. M. (2013). The differences between latent topics in abstracts and citation contexts of citing papers. *JASIST, 64*(3), 627–639.

Liu, X., Yu, S., Janssens, F., Glänzel, W., Moreau, Y., & De Moor, B. (2010). Weighted hybrid clustering by combining text mining and bibliometrics on a large-scale journal database. *JASIST, 61*(6), 1105–1119.

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

Martyn, J. (1964). Bibliographic coupling. *Journal of Documentation, 20*(4), 236.

Mc Cain, K. W. (1983). The author co-citation structure of macroeconomics. *Scientometrics, 5*(5), 277–289.

McCain, K.W. (1989). Descriptor and citation retrieval in the medical behavioral sciences literature: Retrieval over-laps and novelty distribution. *JASIS*, 40(2), 110–114.

Morris, S. A., Yen, G., Wu, Z., & Asnake, B. (2003). Time line visualization of research fronts. *JASIST, 54*(5), 413–422.

Mullins, N. C., Hargens, L. L., Hecht, P. K., & Kick, E. L. (1977). The group structure of co-citation clusters: A comparative study. *American Sociological Review, 42*, 552–562.

Mutschke, P., & Quan-Haase, A. (2001). Collaboration and cognitive structures in social science research fields: Towards socio-cognitive analysis in information systems. *Scientometrics, 52*(3), 487–502.

Mutschke, P., Mayr, P., Schaer, P., & Sure, Y. (2011). Science models as value-added services for scholarly information systems. *Scientometrics, 89*, 349–364.

Narin, F., Pinski, G., & Gee, H. H. (1976). Structure of the biomedical literature. *Journal of the American Society for Information Science, 27*(1), 25–45.

Narin, F., & Noma, E. (1985). Is technology becoming science? *Scientometrics, 7*(3), 369–381.

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

Palacios-Huerta, I., & Volij, O. (2004). The measurement of intellectual influence. *Econometrica, 72*(3), 963–977.

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

Papadimitriou, C., Raghavan, P., Tamaki H. & Vempala S. (1998). Latent semantic indexing: A probabilistic analysis, PODS *Proceedings* of the 17th ACM SIGACT-SIGMOD-SIGART symposium on principles of databases systems. 159–168.

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

Polanco, X., Grivel, L. & Royauté, J. (1995). How to do things with terms in informetrics : Terminological variation and stabilization as science watch indicators. In M. Koenig (Ed.), *Proceedings of the 5th ISSI Intl Conference (River Forest IL, June 7-10, 1995)* 435–444: Learned Information, Medford NJ.

Price, D. J. de Solla. (1965). Networks of scientific papers. *Science, 149*(3683), 510–515.

Price, D. J. de Solla. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27*(5), 292–306.

Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. *JASIS, 61*(9), 1871–1887.

Ritchie A., Robertson S. & Teufel S. (2008) Comparing citation context for information retrieval, CIKM'08, *Proceedings* 17th ACM *Conference* on Information and knowledge management 213–222.

Rocchio, J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), *The smart retrieval system: Experiments in automatic document processing* (pp. 313–323). Englewood Cliffs, NJ: Prentice-Hall.

Ross, N. C. M., & Wolfram, D. (2000). End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine. *JASIST, 51*(10), 949–958.

Rosvall, M., & Bergstrom, C. (2008). Maps of information flows reveal structures in complex networks. *PNAS, 105*, 1118.

Roth, C., & Cointet, J. P. (2010). Social and semantic coevolution in Knowledge. *Social Networks, 32*(1), 16–29.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *JASIST, 41*(4), 288–297.

Scharnhorst, A., Börner, K., & van den Besselaar, P. (Eds.). (2012). *Models of science dynamics: Encounters between complexity theory and information sciences (Understanding Complex Systems).* Berlin: Springer.

Small, H. (1973). Co-citation in the scientific literature : A new measure of the relationship between two documents. *JASIS, 24*(4), 265–269.

Small, H. (1980). Co-citation context analysis and the structure of paradigms. *Journal of Documentation, 36*(3), 183–196.

Small, H. (2011). Interpreting maps of science using citation context sentiments: A preliminary investigation. *Scientometrics, 87*(2), 373–388.

Teufel S., Siddharthan A. & Tidhar D. (2006) Automatic classification of citation function, *Proceedings* EMNLP '06 *Proceedings* 2006 *Conference* on Empirical Methods in Natural Language Processing.

van den Besselaar, P., & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics, 68*(3), 377–393.

Waltmann, L., & van Eck, N. (2012). A new methodology for constructing a publication-level classification system of science. *JASIS, 63*(12), 2378–2392.

Watts, C., & Gilbert, N. (2011). Does cumulative advantage affect collective learning in science? *An agent-based simulation, Scientometrics, 89*(1), 437–463.

White, H. D., & Griffith, B. C. (1981). Author co-citation: A literature measure of intellectual structure. *JASIS, 32*(3), 163–172.

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., & Bassecoulard, E. (2006). Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information Processing and Management, 42*(6), 1513–1531.

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.

Zitt, M., Lelu, A., & Bassecoulard, E. (2011). Hybrid citation-word representations in science mapping: Portolan charts of research fields? *JASIST, 62*(1), 19–39. doi:10.1002/asi.21440.

Zitt M., & Small, H. (2008). Modifying the journal impact factor by fractional citation weighting: The audience factor. *JASIST*, *59*(11), 1856–1860.