Intrication between Information Retrieval and Bibliometrics: the case of scientific domain delineation

Michel Zitt

Lereco Lab (retired),
National Institute for Agronomic Research (INRA), Dept SAE2,
Nantes, France
mzitt@numericable.fr

Abstract. This invited paper summarizes a recent synthesis [8] on the topic of scientific domain delineation. See also [5].

Keywords: Scientific Domain Delineation, Information Retrieval, Bibliometrics.

For decades, Bibliometrics and Information Retrieval developed an increasingly close relationship. Information retrieval (namely its application to science and technology) first emphasized data organization, indexing and query systems. Bibliometrics, driven by exploration of science networks (Price) or evaluation of research systems, was boosted by Garfield's "citation index", a co-word which symbolizes the match between IR and bibliometrics. Kessler's work was a beacon of their opening to mapping techniques.

Bibliometric investigation, whether for cognitive or for evaluative purposes, typically targets particular "domains" defined at some scale. The meso-level scope covers sub-disciplines, fields, large research areas; the frontier with the micro-level of research fronts or topics is quite fuzzy. Field delineation is usually understood in terms of publication sets. Delineation requirements may be quite basic or more demanding, depending on the study's ambitions and the type of domain. High-level requirements are associated to large projects on "difficult" fields such as emerging, transversal or complex areas. Operationalization of delineation typically exploits three models, on their own or in combination. All of them rely on quantitative methods in statistics-probability, data analysis (clustering and factor techniques) and graph-network theory.

Model A covers institutional classifications and nomenclatures that stem from the cooperation between scientists, S&T policy experts, librarians and information scientists. Examples: aggregate classification from a few national and international bodies; national or international patent office classifications; classifications from producers of disciplinary or general databases — with various

degrees of disaggregation. A distinction should be made between institutional, national and international schemes of classification. Schemes that are linked to national or institutional evaluation, especially, are politically loaded and even "international" standards may be biased towards the points of view of stakeholders. In this respect, the way in which disciplines are grouped or broken down bears witness to social and national stakes at a given period. Given a bibliometric project, predefined categories sometimes happen to meet the requirements, although this is not generally the case on transversal or emerging areas.

Model B, in contrast with this sedimented knowledge, relies on ad-hoc IR investigations in a particular query system. In some cases, this appears as a refinement brought to existing groupings. For example, whereas some databases offer categories based on journals lists, variants may be established using some expertise — the same goes for patent categories. It is quite common to delineate relatively simple domains in this way, keeping in mind the limitations of Bradfordian ranked lists. A better compromise requires lexical query (or citation) facilities. A domain of science or technology is seldom captured by global terms so that it becomes necessary to combine restricted queries, using the various adaptive techniques of up-to-date IR, in order to reach reasonable precision and recall.

Model C is based on bibliometric mapping and clustering, which makes "invisible colleges" visible through a variety of bibliometric networks (actors, texts, citations, etc.). A domain is then expected to emerge as a delimited area on a map or graph at the proper scale. In practice, the mental expectations of users seldom match a single *deus ex machina* mapping exercise, which inevitably conveys implicit points of view and technical artefacts. Variable adaptive combinations of top-down phases (cutting the domain out of an extended map of science) and/or bottom-up ones (aggregation of themes in low-level maps) will help. Mapping techniques exploit developments in clustering and community detection (graph unfolding: Louvain, Infomap, SLMA...; spectral methods: LSA, pLSA, LDA...) as well as classical clustering and factor analysis.

The modes of **supervision** (by commissioners, bibliometricians, scientists from the domain...) are crucial to validate the findings at various stages. Model A mostly uses "frozen" IR knowledge, which does not spare discussion on its pertinence in the particular case. Model B queries require experts' mental representation of the field, if only to avoid missing significant subareas. Model C has to deal with peculiarities of the mapping methods. Border areas, particularly, need supervision.

Multi-networks. Scientific IR as well as bibliometrics bring into play the various networks that are explicitly or implicitly created by published S&T: actors, papers, texts, citations, classification categories, sometimes funding information, etc. Citation certainly remains the iconic network of bibliometrics, while IR tradition mostly exploited nomenclature-based indexes and lexical terms. The Internet era shuffled the cards, with the rise of webometrics. Quasi-citations (URL linkages) migrated from journal bibliometrics to Google; PageRank then irrigated webometrics and bibliometrics. The cognitive theory of scientific infor-

mation [2] relies on a multinetwork universe. The associated poly-representation of literature offers to combine or confront approaches using different networks — a spontaneous practice in pragmatic bibliometrics — in order to address a wide classe of issues on a particular dataset.

Thus, many producers of scientometric indicators rely on the default SCI[1] or Scopus classes, or ad-hoc lists of journals, for a rough delimitation. Pragmatic mixes improve the process: lexical or nomenclature index queries; supervised lists of prominent authors, or most cited ones, in multistage and adaptive protocols. For example, a high-precision detection of a core literature may be completed by various recall-oriented extensions (query-expansion, network proximity, etc.) with the possible help of science maps. Protocols involving multi-network design are fruitful. Two networks are generally held as the most precise for thematic analysis: lexical terms (words/phrases from controlled or natural language) and paper-level citations; the author networks are also valuable. Citations and words exhibit analogies and also differences: granularity level, statistical properties, unification issues especially in natural language, dynamic capability — more direct for citation. Citation biases are severe in an evaluation context but somewhat less in a mapping context. Lexical analysis is prone to natural language traps, homonymy, synonymy, metaphors, etc. There is huge literature on the properties of the two universes.

Sequential word-citation protocols give a few examples of how themes delimitation and mapping take advantage of those properties. Among those mentioned in the aforementioned synthesis, the sequence that starts with careful core building, using supervised lexical queries, and that goes on with a quasi-automatic citation-based extension, proved efficient ([6] and related works).

Alternatively, one may first compare techniques, by exploring the citation-way and the lexical way in parallel, and then compare and possibly combine the results. We explore this at a cluster delineation level, namely the detection of areas within a given domain (beforehand delimited by hybrid techniques: nanoscience, genomics...). We established that word and citation techniques shape fairly convergent breakdowns but without coincidence, a phenomenon well characterized in cross-maps [7]. Hence, the two approaches are akin but not substitutable, as each one carries its own infometric and sociological point of view.

The "full hybrid" approaches, where citation and text tokens are mixed, differ from the mildly hybrid ones. Even in naive form, hybrid bibliographic coupling outperformed simple b.c. in Boyack and Klavans studies [3]. Flexible integrated approaches account for differences in statistical distributions [1]. Full hybridization privileges the information science perspective and deliberately overviews the sociological dimension of the two networks.

Most classical techniques rely on citation keys and, in the lexical realm, on word n-grams or more sophisticated linguistic analysis. Instead, featureless representation of information such as character n-grams or bit sequences, proved efficient in our experience. Related n-gram or compression distances are used.

[1] SCI, WoS, WoK product from ISI, Thomson, Thomson-Reuters, now Clarivate.

The cost is the black-box effect, losing any direct semantic interpretation. This can be operated field by field (e.g. actors/ all text/ references) or for the whole document. In the latter case, the representation is savagely hybrid.

Delineation ultimately aims at defining a domain in terms of documents, but this objective may be met through indirect paths (rather than inter-articles coupling) and probabilistic model. Citation-in-context studies, following Small's investigations ([4] and many others) changed the granularity of hybrid thinking in bibliometrics. The vizualisation of citation contexts in online engines is now classical. Narrowing the context of particular citations also has various applications in bibliometric studies: section affiliation of references (introduction, methods, findings...); a step further, association or mutual labelling of words and reference within linguistic units (sentences, paragraphs...). From this fine granularity, hybridization techniques expect a strong gain in precision, whatever the aim, historiography, classification, etc. Prior decomposition of an article in smaller lexical units might enhance the convergence of cocitation and coword clustering in the afore-mentioned cross-mapping exercise.

As previously stated, the frontier with micro-level small topic delineation is fuzzy. Meso-scale bibliometric studies, which involve a breakdown into many topics, cannot generally afford detailed supervision at this micro-level, where themes are detected by some automatic data analysis. However, using cross-maps or hybrid designs, may be helpful to look for robust "strong forms" or at least cores in multinetworks, resisting to change of settings and vantage points. Supervision is nevertheless necessary at the key stages, and costly. For both validity, acceptability and involvement of actors, the careful dimensioning and conduct of supervision is a key factor in large studies.

References

An extensive bibliography is found in [8]. Below, just a few indications on our works, and a few classical milestones.

[1] Boyack, K.W., Klavans, R.: Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? Journal of the American Society for Information Science and Technology 61(12), 2389–2404 (2010), doi:10.1002/asi.21419

[2] Ingwersen, P.: Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. Journal of Documentation 52(1), 3–50 (1996), doi:10.1108/eb026960

[3] Janssens, F., Glänzel, W., De Moor, B.: Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis. In: Berkhin, P., Caruana, R., Wu, X., Gaffney, S. (eds.) KDD'07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 360–369. Association for Computing Machinery (2007), doi:10.1145/1281192.1281233

[4] Small, H.: Co-citation context analayses and the structure of paradigms. Journal of Documentation 36(3), 183–196 (1980), doi:10.1108/eb026695

[5] Zitt, M.: Meso-level retrieval: IR-bibliometrics interplay and hybrid citation-words methods in scientific fields delineation. Scientometrics 102(3), 2223–2245 (2015), doi:10.1007/s11192-014-1482-5
[6] Zitt, M., Bassecoulard, E.: Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. Information Processing & Management 42(6), 1513–1531 (2006), doi:10.1016/j.ipm.2006.03.016
[7] Zitt, M., Lelu, A., Bassecoulard, E.: Hybrid citation-word representations in science mapping: Portolan charts of research fields? Journal of the American Society for Information Science and Technology 62(1), 19–39 (2011), doi:10.1002/asi.21440
[8] Zitt, M., Lelu, A., Cadot, M., Cabanac, G.: Bibliometric delineation of scientific fields. In: Glänzel, W., Moed, H.F., Schmoch, U., Thelwall, M. (eds.) Springer Handbook of Science and Technology Indicators, chap. 2, pp. 25–68. Springer, Berlin (2019), doi:10.1007/978-3-030-02511-3_2