Exploiting keywords life-cycle to analyse the dynamics of an emerging field

Philippe P. Breucker, Audrey Baneyx, Aurélie Delemarle, Lionel Villard, Bernard Kahane, Philippe Larédo

Exploiting keywords life-cycle to analyse the dynamics of an emerging field : an experiment in Nanosciences with the Kwords lab

Audrey Baneyx1, Aurélie Delemarle2, Philippe Breucker3, Lionel Villard2, Bernard Kahane2, Philippe Larédo4

1Université Paris-Est, IFRIS, LATTS UMR CNRS 8134
2Université Paris-Est, ESIEE Management, LATTS, IFRIS
3INRA Sens, Université Paris-Est, IFRIS
4Université Paris-Est, ENPC, LATTS, IFRIS, Manchester Business School (MBS, Manchester Institute of Innovation Research)

Research subject and hypothesis

We argue that it is possible to follow the dynamics of a scientific field as it emerges. Being able to do such allows for example, industrial actors to position themselves regarding their competences and public policy makers to support the development of the field. For that matter, we follow traces of its emergence through publications analysis (Bonaccorsi and Thoma, 2005; Bozeman et al., 2007; Mangematin et al, 2008). Data included in large databases and extracted from the WoS brings us the possibility to characterise a scientific field. The use of traditional methods such as journal or co-journals analysis show the main journals of a field and the relationships between them (Hirsch, 2005, van Raan, 2003). The dynamics of journals and the co-citation indicator can illustrate the structuration of a field (Leydesdorff and Rafols, 2009; Park and Leydesdorff, forthcoming, Leydesdorff and Schank, 2008, Leydesdorff, 2007). Co-authors analysis (Callon et al, 1986; White and McCain, 1998) points to the networks of relationships between authors. It has been used to illustrate the internal structure of a field and its central and most powerful actors. Last, co-word analysis used on title and/or abstract points to the science itself showing how elements of a knowledge bases are articulated (van den Besselaar and Heimeriks, 2006). However, we consider that all these traditional methods used to illustrate the organisation of a scientific field can only be used as it is already structured. Indeed, journal analysis is not possible for an emerging field as it does not yet have reference journals (see also Zitt and Bassecoulard, 2006). Co-authors analysis cannot be used either as a community of scientists with star scientists (Darby and Zucker, 2006) does not yet exist. Last, co-word analysis cannot be used as the field is not defined yet which makes it impossible to choose which words should be selected and which should not.

Following Bonaccorsi (2007), we argue that a keywords analysis can illustrate the dynamics of an emerging field. Bonaccorsi showed that an emerging field has a high rate of new keyword introduction. The technological plateform CorTexT (IFRIS[1]) developed a tool named Kwords Lab to explore the keywords of an emerging field. On the contrary to past studies on top of keywords from the abstract or the title, we also take into consideration keywords provided by individual authors as well as journal keywords. If this choice enriches our analysis it also brings heterogeneity which need to be dealt with.

To test this tool, we based our analysis on the emergence of nanosciences: indeed Kahane and Mogoutov (2007) have already defined the field. Defining the outlines of an emerging field is the first difficulty. As their query brings between 1998 and 2006 543 297 publications, we include automatically in our database all keywords (title, author and journal). This brings us with a total of 8 618 664 keywords before treatments. This large amount of data constitutes the second limit of the work. How to deal with several hundred of thousands of words when traditional textual software cannot do so (we tested Calliope[2], Alceste[3] and ReseauLu[4])? It is to exceed these limits that we have developed an automatic tool which allows to classify keywords based on their significance in

[1] Institut francilien « recherche, innovation et société »
[2] http://www.calliope-textmining.com/kl-en.html
[3] http://www.image-zafar.com/index_alceste.htm
[4] http://www.aguidel.com/fr/?sid=6

the emergence phenomenom. It is then easy for the user to select and study some specific keywords. We thus reduce the total number of keywords to be considered. We call this the analysis of keyword life cycle.

Material and method : Kwords Lab tool

The tool is developed in Php and use MySQL as databases. It allows to deal with a large amount of data and is based on statistical calculus, text mining methods, artificial intelligence and natural processing techniques. These permit to characterise heterogeneous keywords and to propose to the user a classification of keywords.

We define a keyword as a word or a set of words, generally a noun or a noun phrase, which passes on the concept important for the author. These keywords will be used to index the scientific paper and to allow thematic research in journal.

The Kwords Lab associates to each keyword all its lexical variations and close forms (for instance we regroup “nanotechnologies”, “nano-technologies”, “Nano-technology” with “nanotechnology”) using, in the first place, the SoundEx algorithm[5] (NARA, 2007), then an algorithm based on Levenshtein distance[6] (Levenshtein, 1965) and then, the QTClust algorithm (Heyer, 1999) to reunite the different forms of each keyword in clusters.

Results of the tool

For the moment, the tool analyses keywords contained in titles, keywords given by authors and keywords given by journals. It identifies automatically four types of keywords (Turenne and Barbier, 2004):

- The first level of analysis make the difference between *persistent and non-persistent keywords*. Persistent keywords are those that are present over a minimum number of years (defined by the scholar) a minimum number of times (defined by the scholar as well in the friendly interface). On the opposite, non-persistent keywords are those that are not present over the selected period a certain number of times.

- The second level of analysis make the difference between *generic and emergent keywords*. Genericity is calculated as a standard deviation with regard to the average of the number of entrances of the keyword. The smallest the standard deviation, the most the keyword is structuring for the field. Emergence on the contrary shows the dynamics of the field as it is based on the rate of growth of keywords.

References

Bozeman B, Laredo P and Mangematin V. 2007. Understanding the emergence and deployment of "nano" S&T. *Research Policy*, 36/6, 807-812

Callon M., Law J. and Rip A., (ed.), 1986, Mapping the Dynamics of Science and Technology : Sociology of Science in the Real World, London, Mac Millan.

Han Park & Loet Leydesdorff, Knowledge linking structures in communication studies using citation analysis among communication journals, Scientometrics (forthcoming)

Heyer LJ, Kruglyak S and Yooseph S. 1999. Exploring expression data: identification and analysis of coexpressed genes. Genome Res. Nov;9(11):1106-15.

[5] Soundex is a phonetic algorithm for indexing names by sound, as pronounced in English. The goal is for homophones to be encoded to the same representation so that they can be matched despite minor differences in spelling.

[6] The Levenshtein distance is a metric for measuring the amount of difference between two sequences (i.e., the so called edit distance). This distance between two strings is given by the minimum number of operations needed to transform one string into the other, where an operation is an insertion, deletion, or substitution of a single character.

Hirsch J.E. 2005. An index to quantify an individual's scientific research output, Proceedings of the National Academy of Sciences of the United States of America, vol 102, n°46, pp. 16569-16572.

Levenshtein I. 1966. Binary codes capable of correcting deletions, insertions and reversals. Doklady Akademii Nauk SSSR 163(4) p845-848, 1965, also Soviet Physics Doklady 10(8) p707-710, Feb.

Leydesdorff L. 2007. Betweenness centrality as an indicator of the interdisciplinarity of scientific journals, Journal of the American Society for Information Science and Technology, vol 58, n°9, pp. 1303-1319.

Leydesdorff L. and Rafols I. 2009. A Global Map of Science Based on the ISI Subject Categories, Journal of the American Society for Information Science and Technology 60(2) 348-362

Leydesdorff L and Schank T. 2008. Dynamic Animations of Journal Maps: Indicators of Structural Change and Interdisciplinary Developments, Journal of the American Society for Information Science and Technology 59(11), 1810-1818

Mangematin et al, Nanotrendchart newsletter, oct 2008 http://www.nanotrendchart.com/

Peters H. P. F. and van Raan A. F. J. ,1991. Structuring scientific activities by co-author analysis: an exercise on a university faculty level, Scientometrics, 1991, vol. 20, no1, pp. 235-255

The Soundex Indexing System. National Archives and Records Administration. 2007-05-30. http://www.archives.gov/genealogy/census/soundex.html. Retrieved 2007-06-07.

Turenne N., Barbier M. 2004. BELUGA : un outil pour l'analyse dynamique des connaissances de la littérature scientifique d'un domaine. Première application au cas des maladies à prions, In Proceedings of Extraction et Gestion de Connaissances, Hébrail G. and Lebart L. eds., Clermont-Ferrand, 2004, France

van den Besselaar P. and Heimeriks G. 2006. Mapping research topics using word-reference co-occurrences: A method and an exploratory case study, Scientometrics, vol 68, n°3, pp. 377-393.

van Raan A.F.J. 2003. The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific developments, Technikfolgenabschätzung, vol 1, n°12, pp. 20-29.

White H., McCain K. 1998. Visualizing a Discipline: an Author Co-citation Analysis of Information Science, 1972-1995. In Journal of the American Society for Information Science 49 (4), pp. 327-355

Zitt M. and Bassecoulard E. 2006. Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences, Information Processing & Management, vol 42, n°6, pp. 1513-1531.

Zucker LG, Darby MR, Furner J, Liu R.C, and Ma H. 2006. Minerva Unbound: Knowledge Stocks, Knowledge Flows and New Knowledge Production. NBER Working Paper No. 12669, November 2006