A bottom-up approach to assess the interdisciplinarity of journals from a multidisciplinary corpus of bibliographical records

Ivana Roche, Dominique Besagni, Claire François, Marianne Hörlesberger, Edgar L Schiebel

▶ **To cite this version:**

Ivana Roche, Dominique Besagni, Claire François, Marianne Hörlesberger, Edgar L Schiebel. A bottom-up approach to assess the interdisciplinarity of journals from a multidisciplinary corpus of bibliographical records. S&TI-ENID 2013, Sep 2013, Berlin, Germany. hal-00960347

HAL Id: hal-00960347
https://hal.science/hal-00960347

Submitted on 18 Mar 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A bottom-up approach to assess the interdisciplinarity of journals from a multidisciplinary corpus of bibliographical records[1]

Ivana Roche*, Dominique Besagni*, Claire François*,
Marianne Hörlesberger**, Edgar Schiebel**

**ivana.roche@inist.fr; dominique.besagni@inist.fr; claire.francois@inist.fr*
CNRS, Institut de l'Information Scientifique et Technique, UPS 76, 2 allée du Parc de Brabois,
Vandœuvre-lès-Nancy, F-54519 Cedex, France
***marianne.hoerlesberger@ait.ac.at; edgar.schiebel@ait.ac.at*
AIT, Austrian Institute of Technology GmbH, Donau-City-Strasse 1, 1220 Vienna, Austria

Abstract
This work is investigating the possibility of assessing the interdisciplinarity of scientific and technological journals without using any taxonomy or classification scheme as those usually adopted in bibliographical databases. For that, we start from a large corpus of bibliographic records from which we extract terms, either keywords already present or obtained by text mining techniques. With the help of a clustering method, that corpus is split into clusters defining a given number of scientific fields. Those fields and the keywords indexing each document are the basis of the calculation of an interdisciplinarity score applying the diffusion model approach. Then, we calculate an interdisciplinarity indicator for each journal by combining the scores obtained by its articles.

Introduction
Interdisciplinarity and its corollary, specificity, are useful characteristics to locate metaphorically a journal in the scientific and technological (S&T) literature landscape. Previous works were done on the issue of subject classification and the creation of coherent journal sets. Usually these approaches are based on data available from Thomson Reuters' Journal Citation Reports (JCR) where the citations got by each publication are aggregated at the journal level, i.e. the work by Leydersdorff and Rafols (2011) that makes use of measures of interdisciplinarity like network indicators or unevenness indicators. In a recent study Thijs *et al.* (2013) introduced a very interesting approach building a network among journals based on bibliographic coupling. Hybrid approaches based on citation-based and lexical similarities are also known (Janssens *et al.*, 2008) as well as approaches combining several indicators of journal specificity based on textual coherence and research communities (Boyack and Klavans, 2011).
In our work, an exclusively content-based approach is developed to determine from a large multidisciplinary corpus of bibliographical records an indicator measuring the interdisciplinarity of each record and, from that, the interdisciplinarity of the journals where these documents were published.
Our approach is directly inspired on the methodology developed in the context of the DBF project (Hörlesberger, 2013), the goal of which was to infer attributes of 'frontier research' in peer-reviewed research proposals under the scheme of the European Research Council (ERC). To this end, indicators across scientific disciplines and in accord with the strategic definition of frontier research by the ERC are elaborated, exploiting textual proposal information and other scientometric data of grant applicants. In particular, an indicator was devised to

[1] This work was partially inspired by DBF (Development and Verification of a Bibliometric Model for the Identification of Frontier Research), a Coordination and Support Action of the IDEAS specific programme of the European Research Council (ERC). The authors wish to acknowledge this contribution.

characterized any project that *"... pursues questions irrespective of established disciplinary boundaries, involves multi-, inter- or trans-disciplinary research that brings together researchers from different disciplinary backgrounds, with different theoretical and conceptual approaches, techniques, methodologies and instrumentation, perhaps even different goals and motivations"* (EC, 2005) and that we defined as interdisciplinarity.
That indicator is built upon the basic assumption and previously successfully tested concept (Schiebel *et al.*, 2010) that the frequency of occurrence and distribution of discipline specific keywords in scientific documents can be used to classify and characterize disciplines. The concept is consistent with the practice of bibliometric clustering, where the contents of each cluster (e.g., words and articles, or cited references and articles) are ranked by some index (e.g. TF-IDF) of specificity to the cluster.
In the next section the developed methodology is presented followed by some preliminary results.

Methodology
Our methodology is based on a diffusion model approach (Schiebel *et al*., 2010; Roche *et al*., 2010), developed in the context of a previous project aiming at detecting emerging technologies, that evaluates the status of each term in a considered discipline by measuring its so-called degree of diffusion.
The diffusion model is founded on the assumption that new findings in a research field are published in journals, conference proceedings, books etc. That S&T literature is collected in bibliographical databases where the content of each document is represented with a set of keywords. Keywords that describe the innovative results occur in the first stage in an unusual manner. In the second stage the research intensifies and established keywords are used. In later stages, the results cross the disciplinary barrier by diffusing to other research fields where they follow a similar evolution cycle. Consequently, the diffusion status is obtained by the calculation for each keyword of a diffusion degree that can be either "unusual", "established" or "cross-section".
Two pragmatic approaches are successively employed to realise this categorisation. Firstly, the so-called Home Technology terms (H-T terms) are defined. We assumed keywords which are specific for a field occurred with a higher probability in that field rather than in others. The probability is defined by the frequency of one term in a field divided by the number of articles in this field, namely the relative term frequency (rtf_{Field}). For a term, we calculate its rtf_{Field} in each field and the field with the highest probability is declared to be its Home Technology. So after this assignment we obtain for each field the list of its H-T terms. Therefore the complete terminology associated to a field consists of the union of its H-T term list and the set of terms imported from the other fields.
Secondly, we use the Gini index (or also Gini coefficient), a measure of statistical dispersion developed by the Italian statistician Corrado Gini (Gini, 1921) at the beginning of the 20th century. The Gini index (*GINI*) is a measure of the inequality of a distribution and it varies from *0* to *1*, a value of *0* expressing total equality and a value of *1* maximal inequality. It is commonly used as a measure of inequality of the income or wealth of the countries. It is, in this study, employed as a measure of the dispersion of a term in a scientific domain. A Gini index equal to *0* means a completely uniform distribution and indicates that the term occurs in all the considered fields of the domain. Conversely, a Gini index of *1* tells us that the term is very specifically limited to the only field where it appears.
If we consider a set of *n* Home Technologies, the Gini index of a term can be calculated by the Brown formula:

$$GINI = 1 - \sum_{k=0}^{n-1} (X_{k+1} - X_k)(Y_{k+1} + Y_k)$$

where *X* is the cumulative share of Home Technologies, and *Y* the cumulative share of occurrences of the considered term.
In the present study, we are not concerned with detecting innovative technologies but with characterizing scientific fields by analysing their related terminologies. So, hereafter we will not speak anymore of Home Technology but of Home Field (H-F). Moreover, we do not consider the categorization of the keywords according to their diffusion degree, but only the determination for each keyword of its rtf_{Field} and its Gini index. These values allow for each keyword either to assign it to a H-F or to discard it if it is not discriminant enough.
Determining the set of H-F can be done either with the help of a pre-defined taxonomy, or with a content analysis approach starting with a term extraction, followed by a validation step operated on the extracted keywords indexing each document and finishing by a clustering splitting the corpus into a given number of H-F.
In our case, we apply a non-hierarchical clustering algorithm, the axial K-means method, coming from the neuronal formalism of Kohonen's self-organizing maps, followed by a principal component analysis in order to represent the obtained clusters on a 2-D map (Lelu and François 1992). This step is realized by employing an in-house software tool, Stanalyst (Polanco *et al.* 2001), dedicated to the scientific and technical information analysis.
The axial K-means is a variant of the well-known K-means clustering algorithm: it derives half-axes, or "axoïds" maximizing a global inter-axes inertia criterion, instead of deriving cluster centroïds maximizing the inter-class inertia. One can sort the cluster's describers and documents along one of these half-axes as well as project the other terms and documents onto it. These projections on any given axis represent the weight of the describers and the document in the corresponding cluster. In this way, one can derive a fuzzy interpretation of the resulting axes, though the method is a strict clustering technique. This method is fast and can handle very large amounts of data. It is formally related to neural models with unsupervised winner-take-all learning. With that clustering algorithm a document may belong to one or more clusters. We consider that the cluster where the document has the higher weight is its H-F.
At this stage, as each publication is allocated to an H-F, it is possible for each one to calculate its HFT and AFT, respectively, the share of its H-F terms and the share of its "abroad terms", i.e. terms assigned to the other H-F. The higher its AFT value, the more interdisciplinary the publication. However this value does not account for the diversity of origins of these imported terms: do they come from a unique H-F or from several? Indeed, for two publications with an equal value of AFT, the one with the greater number of different origins of abroad terms should receive a higher value.
Finally, for each journal represented in the corpus with a statistically significant number of articles, we combine the AFT values calculated for all its articles.

Results

The data set is extracted from the PASCAL database that is specifically adapted to the purpose of our approach. It provides broad multidisciplinary coverage of scientific publications and contains nowadays about 20 million bibliographic records from the analysis of the scientific and technical international literature published predominantly in journals and conference proceedings. On the other hand, the PASCAL records benefit from an indexing by both keywords and thematic categories of a classification scheme assigned to each individual publication, either manually by scientific experts or automatically based on a content analysis. It is this terminology, formed by the indexing keywords that we can also refer to as "terms", that we employ in our analysis, after an assessment step done by a scientific expert.
The query operated in this work aims to represent the whole spectrum of disciplines in the PASCAL database by following the magnitude of their representation. The obtained corpus,

that comprises 105,254 bibliographic records, is a set of randomly chosen weekly updates of the PASCAL database.
Although each PASCAL record has at least one classification code, we did not exploit this information to define our H-F because the fine-grained classification scheme produces a huge number of disciplines, way too large for our purpose. As indicated previously, to determine our set of H-F without the help of that in-house taxonomy, we decided to unfold the "content analysis + indexing validation + clustering" sequence described in the Methodology section.
The examination of the clusters and their content bring to a final validation of the cluster list corresponding to the list of H-F. That is at that time, for instance, that some clusters the content of which is very close could be merged.
At this stage, we have the list of H-F, the list of documents assigned to each H-F and the set of keywords representing the content of each document. So, all the conditions have been met to apply the diffusion model approach to calculate an interdisciplinarity score for each document.
Then, the set of values got by all the documents from the same journal are combined to produce a journal interdisciplinarity indicator. For obvious statistical reasons, only the journals with a significant number of articles in the studied corpus are taken into consideration. Finally, we put the obtained results into perspective.

References

Boyack K.W. and Klavans R. (2011). Multiple dimensions of journal specificity: Why journals can't be assigned to disciplines, Proceedings of the 13th International Conference of the International Society for Scientometrics & Informetrics, Durban, 04-07 July, pp. 123-133

EC – European Commission (2005). Frontier research: The European Challenge. High Level Expert Group Report, EUR 21619

Gini C. (1921). Measurement of inequality of incomes, Economic Journal 31, 124–126

Hörlesberger M., Roche I., Besagni D., Scherngell T., François C., Cuxac P., Schiebel E., Zitt M. & Holste D. (2013). A concept for inferring 'frontier research' in grant proposals, submitted to Scientometrics

Janssens F., Glänzel W. & De Moor B. (2008). A hybrid mappping of information science, Scientometrics, 75, 3, pp. 607-631

Lelu A. & François C. (1992). Hypertext paradigm in the field of information retrieval: A neural approach, 4th ACM Conference on Hypertext, Milano

Leydesdorff L. and Rafols I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations, Journal of Informetrics 5, pp. 87-100, doi:10.1016/j.joi.2010.09.002

Polanco X., François C., Royauté J., Besagni D. & Roche I. (2001). STANALYST: An integrated environment for clustering and mapping analysis on science and technology, 8th International Conference on Scientometrics and Informetrics, July 16-20, 2001, Sydney, Australia, Proceedings Vol. 2, pp. 871-873

Roche I., Besagni D., François C., Hörlesberger M. & Schiebel E. (2010). Identification and characterisation of technological topics in the field of Molecular Biology, Scientometrics, 82, pp. 663-676

Schiebel E., Hörlesberger M., Roche I., François C. & Besagni, D. (2010). An advanced diffusion model to identify emergent research issues: the case of optoelectronic devices, Scientometrics, doi: 10.1007/s11192-009-0137-4

Thijs B., Zhang L. & Glänzel W. (2013). Bibiographic coupling and hierarchical clustering for the validation and improvement of subject-classification schemes, 14th International Conference on Scientometrics and Informetrics, July 15-19, 2013, Vienna