Topic diversity: A discipline scheme-free diversity measurement for journals

Yi Bu buyi@pku.edu.cn, Peking University
Mengyang Li 446218924@qq.com, Peking University
Weiye Gu gwyye@pku.edu.cn, Peking University
Win-Bin Huang huangwb@pku.edu.cn, Peking University

Extended abstract:

In the 17th century, René Descartes created analytic geometry, a brand-new domain that combines both algebra and geometry. This is one of the greatest contributions in mathematics that links "numbers" with "graphs" into a whole. The innovative creativity of analytic geometry is a typical case of interdisciplinarity, a process that bridges multiple disciplines in order to resolve complex issues that previously may not have been easy to address with one single discipline. In recent years, many countries have started to release new science policies to encourage interdisciplinarity. For instance, China founded its 14th division in its National Natural Science Foundation; the new division explicitly focuses on interdisciplinarity, revealing the ambition of motivating methodological, theoretical, and/or cultural "fusion" among established disciplines.

How to quantify interdisciplinarity (e.g., is the interdisciplinarity of publication A greater than that of B)? Information scientists adopt the references of a certain publication as a proxy for characterizing its interdisciplinarity. More specifically, people count the number of disciplines a publication's references occupy—this is called variety. For instance, if publication A refers to prior papers in three disciplines while publication B five, we tend to believe that B is more interdisciplinary than A. A second dimension considers whether disciplines of publications are balanced. For example, given two publications (say C and D) both having 30 references from Physics and Chemistry, if 15 of C's references are from Physics and the other 15 Chemistry, while D has a distribution of 29 and 1, we say that C tends to be more interdisciplinary due to its better balance. Practically, the Gini index, a quite famous measurement to characterize income inequality in economics, is often utilized to quantify balance. Nonetheless, the two dimensions, variety and balance, are insufficient. Scientometricians have proposed a third perspective, namely disparity, to understand semantic differences regarding disciplines. For example, publication E cites references from Physics and Chemistry while publication F references those in Physics and Arts. People may believe that F seems more interdisciplinary because of the disparity between Physics and Arts compared to that between Physics and Chemistry. The "variety-balance-disparity" framework has been a norm in interdisciplinarity studies for a few years.

It may be apparent that, in this framework, "disciplines" of publications/references need to be clearly defined prior to quantification of interdisciplinarity, which requires manually assigned categories in bibliographic datasets, e.g., Web of Science, Microsoft Academic Graph, Scopus, etc. Yet, there are at least three limitations in adopting such human-assigned schemes for diversity measurements:

(1) These schemes are static so they cannot reflect the dynamics of discipline/subject evolution and structures. For example, artificial intelligence is merged with multiple disciplines; thus, a static scheme may not reveal the real-time disparity among disciplines.

(2) Discipline/subject classification is subjective and often fails to gain a consensus from different perspectives. For example, what Information Science "looks like" to one domain expert might be quite different to how another views it.
(3) The granularity of disciplines does not allow more in-depth analyses in, for instance, sub-fields or research topics.

These three drawbacks inspire us to rethink the usage of this three-dimensional framework. To this end, we propose a new diversity measurement, particularly for academic journals, that does not depend on any existing subject classification scheme; this new measurement for measuring academic journals' interdisciplinarity is called topic diversity (TD). This new measurement requires as inputs the abstracts of publications in a certain journal; its output is a real number that quantifies how diverse the research topics of this journal are.

There are a great many details regarding how we calculate the indicator, but there are mainly four steps in the calculation, namely: (1) word extraction, (2) network construction, (3) topic detection, and (4) diversity calculation. In the first and the second steps, we implement some basic natural language processing and select candidate words that are semantically "meaningful" to the following steps by considering the topological structure of the co-word network. In the third step, we detect candidate topics (communities in the co-word network) and filter them to obtain reasonable topics. In the last step, we calculate TD of each journal by considering variety, balance, and disparity.

To verify this indicator, we adopt the Microsoft Academic Graph, a large-scale bibliographic dataset that contains billions of records of publications, authors, citing relations, and other metadata, to investigate differences regarding our indicators and existing ones. We observe that our proposed indicator has a better distinction than existing ones, indicating its feasibility and validity. As a whole, the highlight of this indicator is that it defines and quantifies "disciplines" with some natural language processing and network analysis techniques, rather than relying on existing human-made subject classification systems that may result in some biases.