Measuring Tech Emergence: Selected bibliometric approaches for the identification of emerging and promising technologies

Contribution to the special session "Measuring Tech Emergence" at the 9th Annual Global TechMining Conference, 2019.17.10, Atlanta Georgia

*K*eywords: emergent technologies, novelty, growth, interdisciplinarity, synthetic biology, combinatorial novelty, bibliographic coupling, knowledge growth factor

Introduction:

It is of great interest for decision makers in science, politics and industry to identify emerging topics and technologies. Rotolo et al. (2015) define an emerging technology through the five following key characteristics: radical novelty, relatively fast growth, coherence, prominent impact, and uncertainty and ambiguity. For decades, the bibliometric community has put much effort to identify these emerging technologies with a high potential for future impact in a quantitative way, referring to scientific publications or patent data (Small et al., 2014). There are basically two ways to progress. Most bibliometric studies generally start out with a given small topic which has been identified as emerging and analyze its characteristics. In contrast, a smaller number of publications apply data mining, text mining, citation analysis and bibliographic coupling networks to structure a given technology field with the help of cluster and mapping algorithms and subsequently assess the level of emergence in the found clusters (e.g. Small et al. 2014 and Shibata et al. 2011). As a contribution to the Tech Emerging Contest at the 9th Global Tech Mining Conference and the second strand of publications identifying emerging technologies, we analyze the given technology field "synthetic biology". The contest searches for methods that provide a reproducible procedure to identify emerging R&D topics within a technology domain. The data is gathered from the Web of Science (WoS) database. The challenge is to best predict topics that are notably active in the two last years of research measured by three parameters: (1) scale which addresses the focus within the technology domain, (2) final data set which has to be created based on some basic key words, (3) analytical approaches used to (4) finally provide a ranked list of ten to thirteen precise terms that describe emerging topics in the respective field (output).

Methodology

We extracted multi word phrases and some relevant single word phrases using a general stop word list from the fields title (TI) abstract (AB), authors keywords (DE) and keywords from the ID-field with a transformation from plural to singular. We selected 500 phrases with at least 10 publications.

We used three indicators for the identification of emerging issues: novelty, growth and growing interdisciplinarity. The **novelty** was measured by the age of the knowledge base. Therefore, we extracted the references of all publications of a single or multi word phrase and calculated the weighted mean year of the number of cited references per publication year. The higher the mean year, the younger is the knowledge base and, thus, identified as more emergent. For the **growth** indicator we used two approaches: the weighted mean year of all publications and the quotient of the last two-year periods (number of publications 2014 and 2015 divided by the number of publications 2012 and 2013). **Growing interdisciplinarity** was measured by means of the subject categories (SC) of WoS. We extracted the SC of each publication and counted the number of different SCs per year. The measure for the growth was the mean of relative growth per year (xt+1)/xt for the last 4 years.

For each indicator we calculated the following normalized value: (indicator value – min value)/(max value – min value). The threshold for each indicator was the median of the respective distribution. This procedure delivered fifty phrases. In a last step we selected 13 meaningful phrases for emerging issues and sorted them by descending overall score.

Time series modelling and forecasting for the identified keywords: The evolution of the knowledge base was examined for each keyword using time series modelling techniques, specifically ARIMA modelling. The modelling was performed on relative growth rates of the knowledge base to account for exponential growth patterns (Schiebel and Asenbeck, 2017 and Förster et al, 2018). Thus, the time series models provide a description of the stochastic process generating the sample of observations around the exponential trend. We frequently observed processes with mean-reverting behavior or in some cases white noise. The time series' past behavior was then used to obtain forecasts of the short-term development of the knowledge base related to the keyword.

Results:

The identification of emerging terms delivered the following 13 emerging keywords: adaptive evolution, DNA assembly, metabolic engineering, synthetic circuit, DNA target, protein synthesis, real-time PCR, homologous recombination, target gene, saccharomyces cerevisiae, biosensor, bacillus subtilis, drug delivery.

Discussion:

We found that working only with the identification of keywords is not efficient, because they are taken out of their context and very often multiple terminologies are used to describe research on new technologies. This makes it difficult to judge the emergence of a technology issue. We propose additional approaches. First, staying closely to the analysis of keywords, the so called combinatorial novelty based on the timely development of co-occurrences of keywords is a further step to put keywords pairwise in a relation that gives us more information on technologies than stand-alone keywords. Second, we propose to use bibliographic coupling to identify clusters of similar publications. Taking only publications of the last 2-5 years into account delivers new research fronts of the recent past while the number of publications to be analyzed is reduced. Third, the novelty of each cluster can be measured by the age of the knowledge base. Fourth, the keywords of the publications of the novel clusters can be extracted and assessed with the above-mentioned indicators. Additional data for the measurement and prediction of growth can be collected with a keyword-based search for selected relevant keywords of the clusters. We made a complete analysis for the identification of emerging keywords as mentioned above with the steps (1) to (4). Additionally, we made some first trials with combinatorial novelty, identified research fronts for synthetic biology with bibliographic coupling for the latest two years of the dataset. These approaches and some results will be discussed during the special session

References

Förster, M., Stelzer, B. and Schiebel, E. (2018). "Stochastic analysis of citation time series of emergent research topics", *Proceedings of the 23rd International Conference on Science and Technology Indicators (STI 2018)*, http://sti2018.cwts.nl/proceedings

Rotolo, D., Hicks, D. and Martin, B. (2015). “What Is an Emerging Technology?”, *Research Policy*, 44(10), 1827–43.

Schiebel. E. and Asenbeck, B. (2017). “The Knowledge Growth Factor KGF as a new indicator for the quantification of the emergence of research issues - The case of tribological wear”, *Atlanta Conference on Science and Innovation Policy*, 177–179.

Shibata, N., Kajikawa, Y., Takeda, Y. and Matsushima, K. (2008). “Detecting emerging research fronts based on topological measures in citation networks of scientific publications”, *Technovation*, 28(11), 758–75.

Small, H., Boyack, K.W. and Klavans, R. (2014). “Identifying emerging topics in science and technology”, *Research Policy*, 43(8), 1450–67.