Big-data-augmented approach to emerging technologies identification: case of agriculture and food sector

Ilya Kuzminov [0000-0002-7321-3934], ✉Pavel Bakhtin [0000-0002-0572-3983], Alina Lavrynenko [0000-0001-5311-9898]

National Research University Higher School of Economics, Moscow, Russian Federation

ikuzminov@hse.ru
pbakhtin@hse.ru
alavrinenko@hse.ru

Abstract. The paper discloses a new approach to emerging technologies identification, which strongly relies on capacity of big data analysis, namely text mining augmented by syntactic analysis techniques. The opportunities of the new big-data-augmented methodology are shown in comparison to existing results, both globally and in Russia. The integrated ontology of currently emerging technologies in A&F sector is introduced. The directions and possible criteria of further enhancement and refinement of proposed methodology are contemplated.

Keywords: Text Mining · Emerging Technology · Agriculture and Food Sector.

1 Introduction

Technology identification and mapping exercises for effective science and technology (S&T) and innovation policies shaping become less feasible without modern data science techniques application. This happens due to the explosive growth of diversity and quantity of available S&T information, drawbacks of human-performed analytics, as well as overextended periods of foresight studies and budget limitations. The attempts to solve the problem include tech mining [1–3], as well as creation and regular update of the ontologies specific for foresight studies [4]. The main disadvantages of these approaches are their insufficient scalability, as well as strong reliance on large expert validation, manual filtering and data outputs cleaning. The results are highly prone to subjectivity, human errors and obsolescence.

For the purposes of emerging technologies (new technologies that might have a significant impact on the economic activity) identification, we see text mining / semantic analysis tools as the most appropriate, as identification of new man-made phenomena of known nature (technologies in this case) can be reduced to identification of new syntactic constructions signifying them. The fact that man-made artifacts tend to be explicitly named, described and discussed with the use of written language makes the problem well-posed.

To demonstrate text-mining-augmented techniques applied to technology identification and mapping we consider the case of the agriculture and food (A&F sector). Our choice is driven by the fact that large proportion of global challenges are directly related to A&F sector [5], and seemingly cannot be solved without radical technology innovation across the globe [6].

2 Methodology

The main hypothesis in this paper is that "emerging technology" as a signifying syntactic construction has not lost its semantic utility despite the hype around this concept. The analysis is based on the ample material of the two-year A&F sector foresight study, and relies on the capabilities of the Text Mining System of the National Research University Higher School of Economics (NRU HSE). Composition of data sources of the system include stratified random sample of summaries and metadata of top cited research papers and international patents, as well as newsfeeds items from tops of global news portals with science and technology flavor, analytical and forecast reports, declarations, proceedings and other documents in PDF format (all acquired through open access sources). At the time of the study, the system featured more than 12 million documents, several hundred million sentences, of which up to 3 million documents were at least partially relevant to A&F sector and adjoining sectors, such as biotechnology and bioenergy, more than 150 million terms - object signifiers (among which technologies are presented).

In this paper, we present one of possible approaches of technology identification, namely cascade identification of words being governors within terms. The method allows identifying unigrams – universal signifiers of semantic field of "techologicality", i.e. words that radically increase the probability of an n-gram containing them to be a name of certain technology. Examples of such words are technology, method, system, platform, model, tool, layer, enzyme and others. Extraction of all object-signifying words allows getting hundreds of thousands of terms – candidates for being names of technologies (for instance, DNA sequencing technology, or recirculating aquaculture system, etc.). These lists are filtered with the use of author-built machine learning algorithms dealing with "information-richness" of terms, their monopolism and specificity and other attributes.

Then, analysis of dynamics of presence intensity in the discourse during the last years is conducted for the candidate technology-signifying terms. The relative frequency of terms is calculated using the following formula [7]:

$$f_{term} = \frac{\sum_{i=1}^{n=amount\ of\ documents} s_{term,i)}}{s} \quad (1)$$

where i – document's number, $s_{term,i}$ – the amount of sentences in the i-th document, in which a term has occurred, s – the amount of all sentences in the corpus.

In order to calculate the dynamics of the candidate technology-signifying terms, we adapted the formula of average annual growth rate (AGR):

$$AAGR_{term} = \frac{\sum_{i=second\ year}^{m=last\ year}(\frac{f\,term_i}{f\,term_{i-1}}-1)}{n-1} \quad (2)$$

where n – the amount of years, for which the collection of documents is available, f_{term_i} – relative frequency of term in i-th year.

The results interpretation rest on the assumption that currently unfolding technology trends (including the development and adoption of emerging technologies) are characterized by the growth of interest towards them at least in one of the corpora of documents (science, patents, news and blogs, analytical reports). It is suggested that emerging technologies with strong potential of surviving and upscaling to the global production systems have a signature of ever-increasing public awareness of them.

3 Findings

A&F technologies identification results include the list of 181 items. A random sample is provided below:

- aeration technologies
- agricultural conservation technologies
- agricultural drones
- algal biofuel technologies
- bioconversion technologies
- cultured meat technologies
- dairy technologies
- DNA micro array technologies
- feed probiotics
- fertilisation technologies
- horticultural technologies
- integrated soil fertility management technologies
- LEISA technologies
- meat processing technologies
- smart irrigation

Technologies were distinguished by dynamics of intensity of their presence in the discourse during the last years. It can be visualized as trend maps: 2-dimensional plots with one axis representing the popularity of a term and the other showing the year-by-year dynamics of the normalized popularity (relative frequency of use). For trend map of technologies in agriculture on media resources see Fig.1. The upper-right quadrant consists of the strongest topics shaping the future agenda of the sector, they are popular and gaining traction: in media they are exemplified by CRISPR technologies, agroforestry and aquaponic technologies, precision agriculture and microalgae technologies etc. The lower-right quadrant contains the so-called "weak signals": they are highly trending but underrepresented in discourse yet. They can contain the emerging technologies. This group presented by smart irrigation technologies, molecular breeding and zinc-finger nucleases technologies etc. Among the popular topics losing their significance are fertilisation, pruning, antifouling technologies and many more.

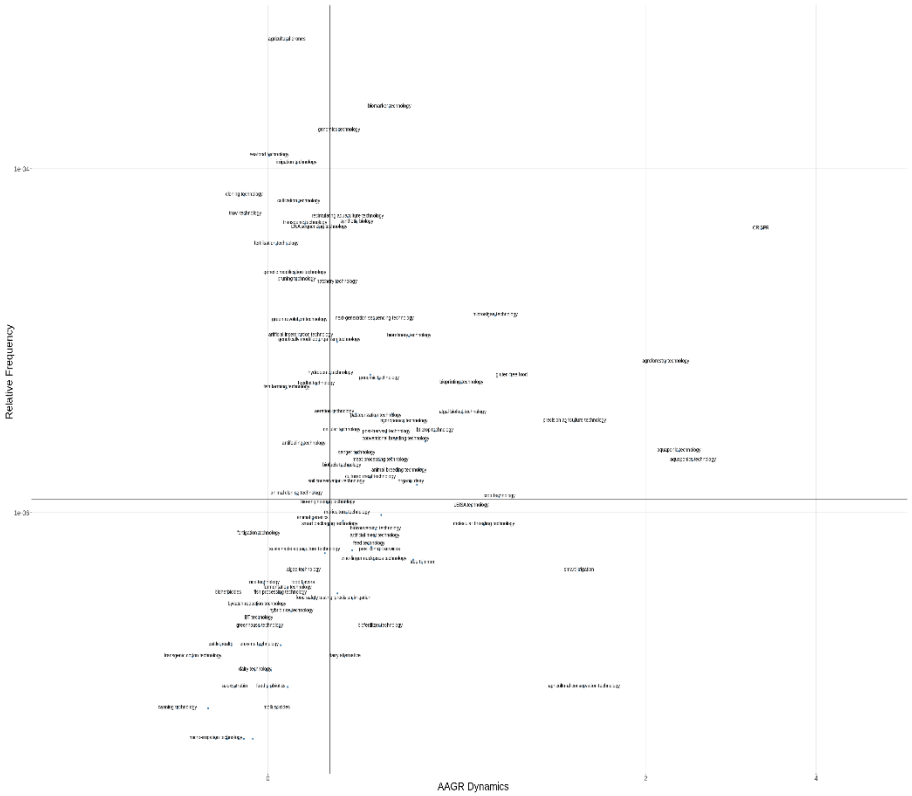

Fig. 1 Trend map of agricultural technologies on media resources.

4 Discussion and Conclusions

The method applied in this study yields most results with data obtained on research papers and especially patents abstracts, which contain less low-informative terms than general reports of international organizations, and discuss technologies in more concrete terms. Within this approach, any individual terms are filtered out without an anchor term within. One of the consequences is that branded, trademarked and other proprietary technologies are almost not present in the output (with some exceptions, such as Round-Up pesticide, which name has gone almost denominative in the GMO application discourse, so that "roundup technology" were mentioned in the texts analyzed). The next steps of filtering the obtained lists of technologies may include building the semantic map that demonstrates dynamic classification, trend maps based on other sources of data, as well as hype maps that show difference in normalized popularity of topics in different data sources (e.g. media vs patents).

The limitations of such approach is high dependence on the marker terms. Some technologies may never co-occur with “technologicality” terms meaning the algorithm will miss them. In order to overcome this obstacle, future studies will concentrate on two main points: searching terms that are relatively more specific to patent literature

compared to other sources of data (such as scientific publications, media news, analytical reports) as potentially technical terms, as well as using identified technology terms as a sample for machine learning based on word embeddings. In other words, the main hypothesis for the future studies is that terms that are semantically highly similar to technology terms (based on word2vec, GloVe or other approaches) are also likely to be candidates for being names of technologies.

References

1. Porter, Alan L., and Scott W. Cunningham. Tech mining: exploiting new technologies for competitive advantage. Vol. 29. John Wiley & Sons (2004).
2. Madani, F. 'Technology Mining' bibliometrics analysis: applying network analysis and cluster analysis. Scientometrics 105(1), 323–335 (2015).
3. Bakhtin, P., and Saritas, O. Tech Mining for Emerging STI Trends Through Dynamic Term Clustering and Semantic Analysis: The Case of Photonics. Anticipating Future Innovation Pathways Through Large Data Analysis, 341–360. Springer International Publishing (2016).
4. Popper, R. iKNOW project, http://www.foresight-platform.eu/wp-content/uploads/2010/06/5.4-Popper_iKNOW_EFP_final.pdf, last accessed 2017/03/15.
5. Godfray, H., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence., Muir., Pretty, J. Robinson, S., Thomas, S., Toulmin, C. Food security: the challenge of feeding 9 billion people, Science Express, 327(5967), 812–818 (2010).
6. Royal Society. Reaping the benefits: Science and the sustainable intensification of global agriculture, RS Policy Document 11/09, The Royal Society, London (2009).
7. Bakhtin, P., Saritas, O., Chulok, A., Kuzminov, I., Timofeev, A. Trend monitoring for linking science and strategy. Scientometrics, 1-17 (2017)