Towards a Technology Convergence Index for Information Technologies: A Keyword Extraction Approach Applied to ArXiv

Jacques Roitel (jacques.roitel@epfl.ch), Section of Mathematics, EPFL
Dimitri Percia David (dimitri.perciadavid@unige.ch) Information Science Institute, Geneva School of Economics and Management, University of Geneva
Alain Mermoud (alain.mermoud@ar.admin.ch) Cyber-Defence Campus, Armasuisse Science and Technology
Thomas Maillart (thomas.maillart@unige.ch) Information Science Institute, Geneva School of Economics and Management, University of Geneva
Alessandro Tavazzi (alessandro.tavazzi@epfl.ch), Armasuisse

Background

The accelerating pace of technological development is continuously redefining information technologies [1]. In such an environment, a myriad of opportunities emerges to enhance the efficacy and efficiency of operations for all types of social organizations [2]. However, in such a fast-paced and complex context of technological development, opportunities are undeniably also accompanied by threats [3, 4, 5]. Therefore, identifying emerging technologies is paramount for anticipating opportunities and risks in organizations early.

Problem Statement

Extant literature emphasizes that innovation – materialized by emerging technologies – often results from a combination of existing technologies [6]. This phenomenon is coined *technological convergence*, and is defined as 'a breakthrough which combines at least two or more existing technologies into hybrid technologies' [6]. It is mostly studied with patent data [7]. However, patents are registered at a late stage of technological development, which hinders early anticipation due to the intrinsic lagged aspect of such patents with respect to the advent of emerging technologies. Additionally, studies focused on technological convergence often investigate convergence without quantifying it for bench-marking purposes (such as does any index).

Approach

In this work, we address these two concerns by (i) analyzing a type of document – pre-prints of scientific works – that arrive at an earlier stage of technological development and (ii) computing and forecasting a technological convergence index.

Data

We extract data from open scientific works (*i.e.*, scholarly articles consisting of working papers, preprints, technical reports, post-proceedings, and publications) labeled e-prints and uploaded on the arXiv repository. This latter is a free distribution service and open-access archive for academic articles related to various technical fields, including computer science (uploaded e-prints are not peer-reviewed). First, we download the entire arXiv repository (1,858,293 files, corresponding to 3 TB of text in pdf) through a mirror of the database found on kaggle. The data encompasses all e-prints uploaded since the inception of the arXiv repository (August 14, 1991) until December 31, 2020.

To consistently classify, archive, and relate all e-prints to specific technology categories, arXiv representatives (composed of a scientific advisory board) developed a systematic category taxonomy. They determine this taxonomy with a Delphi-like method involving expert members for each arXiv scientific field. This implies that authors willing to upload their e-prints on arXiv must select the corresponding category.

Then, arXiv moderators check the authors' classification to ensure consistency. We consider this three-step classification to be robust as (i) the taxonomy is created through a consensus reached by a panel of experts, (ii) authors have no apparent incentive to misclassify their work, and (iii) moderators check the classification consistency. As e-prints are attached to various predetermined arXiv fields unrelated to computer science (such as physics, mathematics, quantitative biology, quantitative finance, and economics), we filter the arXiv predetermined fields to extract the computer-science technologies (denoted cs.) repository.

For the sake of brevity, we specifically study the 'cryptography and security' subsection and its relation with other subsections.

Methods

The question that arises is the following: how to use arXiv data to model, compute, visualize and forecast convergence between technologies— We hypothesize that two technologies are more likely to converge if they share common topics, that is, if they use similar semantics. Conversely, two technologies that do not share common semantics have few chances of converging.

As stated above, arXiv is composed of categories, themselves composed of subsections. Our study focuses on the computer-science category, and each sub-section is considered a technology. To investigate the relationships between sections, we leverage *keyBERT* [8] for extracting the most relevant and representative keywords of titles and abstracts of e-prints, and this for each subsection over time. The semantics proximity of the 'cryptography and security' subsection with the 'general literature' subsection serves as a baseline (indicating a negligible convergence index, as shown in Figure 1). We then compare these shared keywords' evolution among two technologies to create a dynamic technological convergence index. The advantage of creating such a dynamic technological convergence index from keywords is twofold: (i) the index makes it possible to quantify the technological connection through time, and (ii) the keywords make it possible to dynamically determine subjects on which convergences are evolving – either positively or negatively.

Results

We observe three relevant processes, as shown in Figure 1: technologies related to the 'cryptography and security' subsection are (i) diverging with technologies related to 'information theory', (ii) stagnating with technologies related to 'databases', and (iii) converging with technologies related to 'machine learning' and 'sound'.

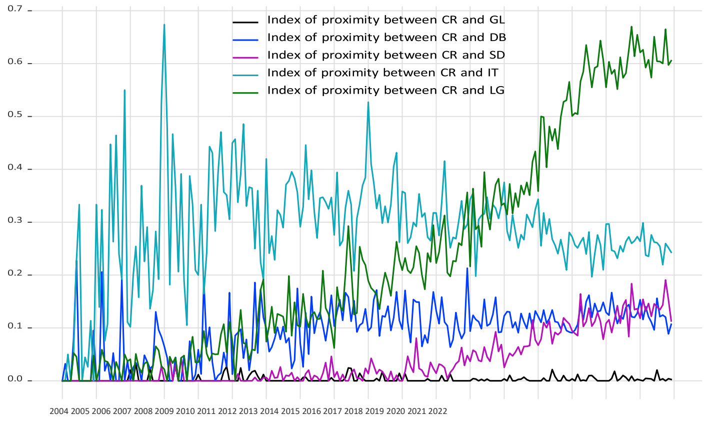

Figure 1: Evolution of indices between 5 subsections and Cryptography and Security between 2004 and 2022. **CR** stands up for 'Cryptography and Security', **GL** for 'General Literature', **DB** for 'Databases', **SD** for 'Sound', **IT** for 'Information Theory', and **LG** for 'Machine Learning'.

Technological convergence indices linking the 40 computer-science subsections (two by two) of arXiv are available upon request. These indices are presented in the form of monthly time series and are included between 0 and 1.

Conclusion

Our contribution is threefold: (i) we create an index that quantifies the phenomenon of technological convergence, (ii) we use pre-prints of scientific work to create such an index, eluding the lagged problem of patents, and (iii) we assumed that technological convergence could be identify through a common vocabulary between two technologies. Next, we apply transfer learning methods to forecast our index (available upon request). Altogether, our work extends the TechMining literature by offering a novel method of informing decision-makers about early technological convergence.

References

[1] John Shalf. "The future of computing beyond Moore's law". In: *Philosophical Transactions of the Royal Society A* 378.2166 (2020), pp. 61–76.

[2] Gerald W. Brock. *The second information revolution*. Harvard University Press, 2021.

[3] Stefan Laube and Rainer B¨ohme. "Strategic aspects of cyber risk information sharing". In: *ACM Computing Surveys (CSUR)* 50.5 (2017), pp. 1–36.

[4] Ross Anderson and Tyler Moore. "The economics of information security". In: *Science* 314.5799 (2006), pp. 610–613.

[5] Julian Jang-Jaccard and Surya Nepal. "A survey of emerging threats in cybersecurity". In: *Journal of Computer and System Sciences* 80.5 (2014), pp. 973–993.

[6] Clive-Steven Curran and Jens Leker. "Patent indicators for monitoring convergence–examples from NFF and ICT". In: *Technological Forecasting and Social Change* 78.2 (2011), pp. 256–273.

[7] T.U. Daim and H. Yal¸cin. *Digital transformations: new tools and methods for mining technological intelligence*. Edward Elgar Publishing, 2022. isbn: 978-1-78990-862-6.

[8] Ian Tenney, Dipanjan Das, and Ellie Pavlick. "BERT rediscovers the classical NLP pipeline". In: *arXiv preprint arXiv:1905.05950* (2019).