Research Article

A Method to Reduce False Positives in a Patent Query

Johannes Van Der Pol[1], Jean-Paul Rameshkoumar[2]

1. Innovation Studies, Copernicus Institute of Sustainable Development, Utrecht University, Netherlands; 2. University of Bordeaux, Bordeaux, France

The aim of this paper is to present a method that allows researchers and analysts to reduce the number of false positives in a patent query. Patents are not only used for prior art searches but increasingly for competitive analyses and the analysis of the evolution of technology. When these cases focus on specific technological domains, non-experts will aim to identify patents related to their focus technology. In certain cases, this can require complex queries containing thousands of patents. It then becomes difficult to identify false positives. We present a method that allows researchers and analysts to refine their queries on large datasets.

Corresponding author: Johannes Van Der Pol, j.p.g.vanderpol@uu.nl

JEL: C80; 031; 034

1. Context and motivation

Patents are increasingly used as a data source for analyses that go beyond prior art searches. In Economics alone, patents have been used since the 1980s for measuring R&D output (Grabowski and Mueller (1972); Jaffe, Trajtenberg, and Henderson (1992)). Since then, patents have been used for a variety of purposes related to the understanding of the evolution of technologies (Saint-Jean, Arfaoui, Brouillat, and Virapin (2020); van der Pol and Rameshkoumar (2018)), for competitive technological intelligence (Coates et al. (2001); Flamand (2016)), for measuring science-industry interactions (Han and Magee (2018); Tijssen, Yegros-Yegros, and Winnink (2016)), and more largely, for the analysis of technological innovation systems (Frigant and Talbot (2005)). These analyses use patent data to identify citations between technological domains and firms, collaborations, the emergence of technological concepts, inventor collaboration, patent transfers, and so on (Ernst (2003); Trippe (2003)).

The quality of the results of these analyses is highly dependent upon the quality of the patent dataset that is used. The analysis of a technological domain by non-experts implies the building of a query to find relevant patents without specific knowledge of the technology. Any false positive can result in incoherent results in terms of citations, collaborations, and textual elements, which we want to avoid since they can lead to false interpretations and hence result in bad decisions. This makes the query a vital piece of the work, even though there is no patent query that can ensure all relevant patents will be retrieved (Trajtenberg (1987)).

For these reasons, and due to the ever-increasing number of patents, it is important to be able to quickly and efficiently identify false positives in a patent query. This issue is different from the patent retrieval issue that has been largely documented (Khode and Jambhorkar (2017); Shalaby and Zadrozny (2019)). We are not concerned with the identification of patents close to a given patent, but rather with patents relevant to a technological domain. In some cases, the construction of a query is a simple task – for instance, if the domain one wishes to analyse is defined by one specific patent classification. Often, however, this is not the case. 5G, structural composite materials for aeronautics, lithium-ion batteries for cars, 3D skin printing, and green tyres are some striking examples. The technologies do not have one specific classification and therefore require combinations of inclusions/exclusions of both different classifications and keywords. In such a case, there is a high risk of false positives due to homonyms, acronyms, bilingual homophones, paraphrases, and synonyms. It would be presumptuous for anyone to affirm that they know all about a technology, and even experts are often surprised by the applications of a given technology. The difficult part of cleaning a patent query is to identify which patents are really out of scope and which are the beginning of a new application or trajectory. One could suggest simply looking at the classifications and excluding anything that does not make sense. However, excluding a classification can result in excluding relevant patents. Bamboo and tires do not appear to have much in common, and yet bamboo fibres can be used in tires. Excluding bamboo might remove certain false positives but will also result in the removal of true positives.

Whether it is for the purpose of understanding the economics of innovation behind these technologies or the industrial dynamics of the strategic behaviour of a firm, analysts and researchers need to be able to identify relevant patents. Whether one has a technical understanding of the technology or not, queries can bring in false positives for a variety of reasons which we will discuss further in this paper.

We will show in this paper how we use classification networks to assess the coherence of a query, identify what is removed when we exclude a classification, and how we quickly identify the classifications to verify. The aim of this paper is to provide analysts and researchers with a method to reduce false positives even without knowledge of the technology.

This paper is organised as follows: we will start by identifying how false positives emerge. Using this information, we will explain how we use classification networks to identify problems in a query. We will show

an example of the method before concluding the paper.

2. How false positives emerge in patent queries

When building a query for complex technologies, it is common to combine classifications with keywords. The use of keywords is common, especially when classifications are too broad or nonexistent. There are no classifications for 5G technologies. To identify 5G patents, we would combine different keywords describing the underlying technologies with classifications on telecommunication. If we search for patents on silica-reinforced rubber, we would combine the classification for rubber with different keywords for silica. However powerful, keywords can bring in a lot of false positives. Suppose we would like to create a dataset containing all patents related to ”carbon.” We would search for the keyword ”Carbon” in the text of the patent. This would bring in many relevant patents, but consider the following patent (US20050150283A1), from which the description reads:

> *”FIG. 11 shows an embodiment in which lines 140 form a diamond-shaped network 141, which is connected at node 142 to transponder 136. [0085] Fiber-like lines 140 are advantageously made of steel cable, carbon, electrically conducting plastic, and other electrical conductors known from aeronautics, for example, and combined with other materials or fibers, e.g. carbon, aramide, steel cable plastic, electrically conducting plastic ceramic fiber, etc.*”

Terms such as ”carbon” and ”aramide” can be keywords used for certain queries, but it is clear from this text that the patent itself is not related directly to these technologies. This problem occurs often when building patent queries using keywords and is unavoidable. Other reasons for false positives/negatives are related to the terms themselves:

- Synonyms: A query should include all synonyms of a term. If not included, some patents might be missed (false negatives/silence). *e.g* tire/tyre
- Homonyms: Result in capturing information that is irrelevant. *e.g* if aiming to find patents related to trains (the transport vehicle), using the term ”train” can capture patents containing variations of the verb ”to train,” resulting in false positives/noise.
- Bilingual homophones: Searching in different languages is a problem and can bring in irrelevant patent documents (soy (the bean) and soy (the verb ”to be” in Spanish), tire (for a car) and tire (”to pull” in French)), resulting in false positives/noise.
- Paraphrase: This problem is especially present in patents. Patent authors will aim to be as vague as possible in their patents. This can result in authors not using certain terms (camera = a tool for taking pictures). This can result in missing relevant information (false negatives/silence).

- Acronyms: Search engines do not always take caps into account; therefore, an acronym can be confused with a word. For instance, positron emission tomography (PET) will be confused with the word pet.

It is often complicated to find a way to remove the false positives without knowing if one does not remove any true positives (through the exclusion of a classification, for instance). Even for an expert who knows the right keywords, it might be complicated to know whether something that is being excluded could not be a true positive. In the method we propose here, we allow for false positives in the first stage. We then analyse the dataset to identify what can be excluded without risking the removal of true positives.

3. Classification Networks for query assessment

When we consider that innovation is achieved by the combination of existing knowledge (Schumpeter (1942), Nelson and Winter (1982)), a combination of IPCs reflects a compatibility or a technological proximity. This means that patents related to a technological domain are somehow connected since they use similar underlying knowledge. We use classifications as proxies for these pieces of knowledge. When a patent query is supposed to represent a technological domain, we can use a classification network for the validation of the coherence of a patent query; we use a classification network. False positives that come from keywords still contain classifications.

By analysing how classifications are related, it will be easier to assess when a patent is completely out of scope or related to the core of the query and valid for the query. Classifications that are not at all connected to the technology are very likely not to be connected to any of the valid classifications. A network will show this immediately, and visually. The core of the method we present hence relies on creating a network of classifications from which we can deduce if certain classifications can be removed.

3.1. Creating a classification network

A classification network is built from the classifications present on the patents of the dataset (any classification will work). Whenever two or more classifications are present on a patent, we connect these classifications. In Figure 1, we show how the classifications (in this case, IPC) are used to create a network. This process is repeated for all patents in the dataset, as shown in Figure 2, creating a large network. Whenever there is a classification in common between patents, they will connect.

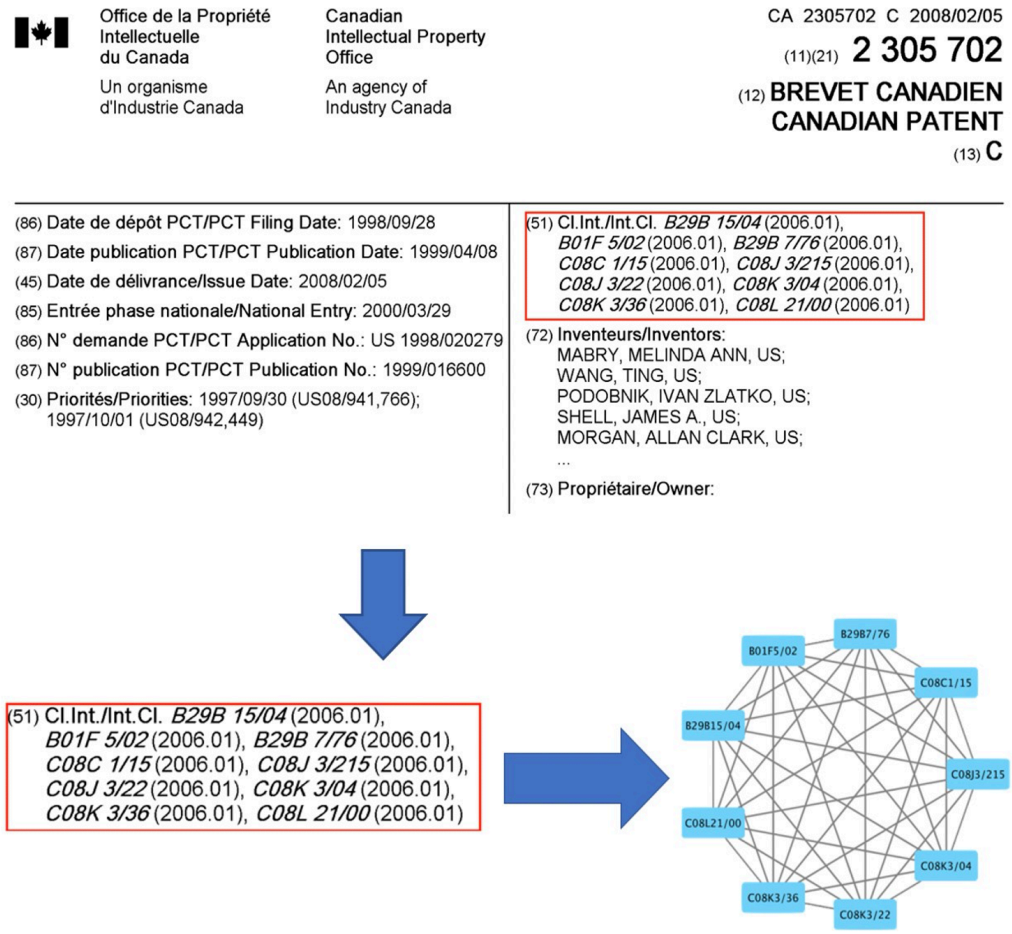

Figure 1. The IPC network is built by connecting IPC classes present on the same patent. This is done for all patents in a given portfolio.

Figure 3 shows a classification network for a technological domain related to tyres; we will use this as an illustration. In this network of classifications, we can see that there is heterogeneity between the classifications in terms of the number of connections (as there should be). Some classifications are more central to the network, some are more at the periphery, and others are not connected at all (components on the top right of the figure). The structure of the networks provides us with insight into the technology we are analysing, showing those classes that are part of the core of the technology and those that are further away (these can be applications). There are classifications that make up the core (C08K and C08L, for instance) and are clearly at the heart of the technology. Some classifications are related to this core, but they are further away. In terms of knowledge, this would imply that they are either applications of the core technology or false positives that are related to the core but out of scope nonetheless. van der Pol and Rameshkoumar (2018) have shown through a

doi.org/10.32388/Y2BOXI

dynamic analysis how this type of network forms over time. They showed that an IPC network emerges with a technological core. Applications arrive later, connecting to the core and changing its structure.

Patent 1	Patent 2
C09C-001/30	B24B-037/00
C08G-077/38	B82Y-030/00
G03G-009/097	C09C-001/30

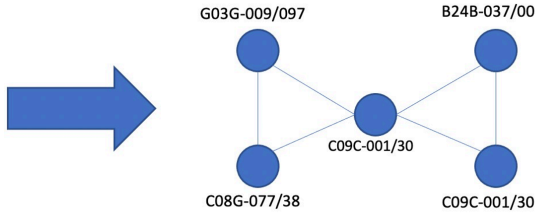

Figure 2. How the classifications of different patents are combined to create a network.

In any case, it stands to reason that all the classifications should be connected, at least loosely.

In order to identify false positives from such a network, we will use a typical three-stage analysis of this network: analysis of the components, the communities, and finally the nodes.

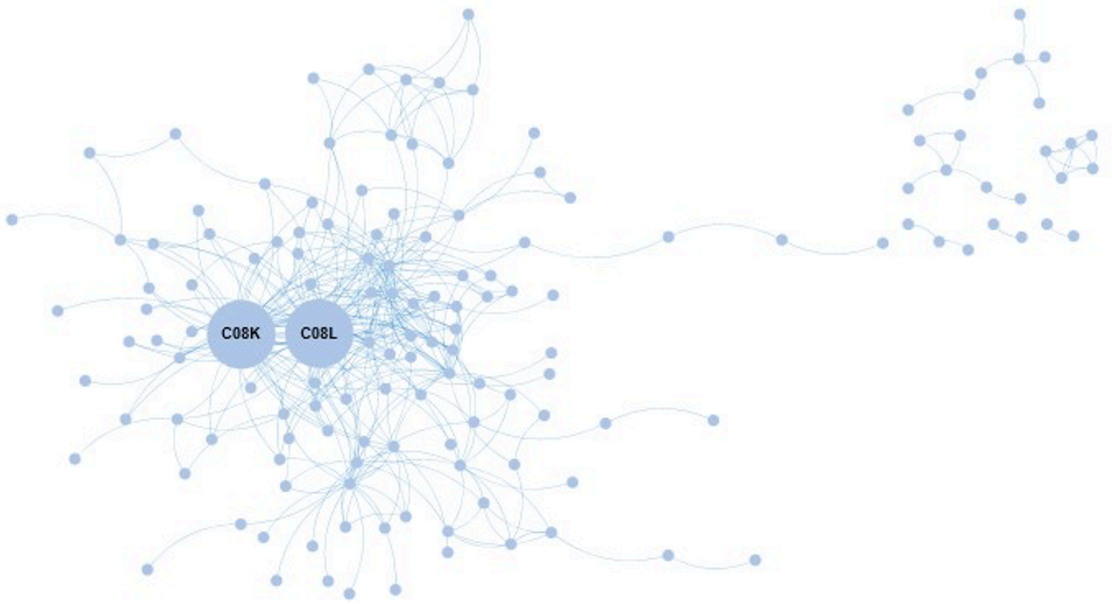

Figure 3. Example of a network of IPC codes for a domain-level patent query.

Source: Questel Orbit, network generated automatically with Intellixir.

doi.org/10.32388/Y2BOXI

4. Structural analysis for the identification of false positives

A normal network analysis consists of three stages: an analysis of components, an analysis of clusters (or communities), and an analysis of the nodes (van der Pol et al. (2018)). The idea behind these steps is to explain how the network is built and identify the underlying rules at work. In our case, we focus on a network that reflects innovation since it builds by a combination of classifications. This means that we want to aim to understand how different classifications connect to result in the larger technological domain. The overall structure of the network reflects the modularity of the technology. A sparsely connected network shows that different parts of the technology are connected to create a whole, while a densely connected network shows a much more automatised technology, where distinctions into modules are less clear-cut. The communities in these networks show closely connected nodes that make up part of the technology; in other words, they are coherent together.

4.1. Components

A component is a part of the network that is not connected to the rest of the network (Barabasi (2013)). In our case, figure 3 has eight components: the seven components on the top right and the large component next to it. These components contain knowledge that is not related to the core of the patent set. This does not automatically mean that the patents with these classifications should be removed. It is possible that these components contain knowledge that is relevant to the core but has not yet connected, but will in the future. The network allows us to quickly identify the classifications we need to check. Our experience shows that 95% of the time, components reflect a mistake in the query.

If a mistake is identified, one can simply remove the classifications from the query by excluding the classification. Since there is no link to other relevant classifications, we know that exclusion will not result in the removal of true positives.

4.2. Communities

A second step in the analysis is the identification of communities inside the network. This step aims to segment the network into communities that represent a specific aspect of the technology (this could be a specific application of the technology, a subdomain, older technology, etc.). We use this segmentation to identify groups of patents that combine different classifications that are or are not at the core of the set we aim to build. Different techniques exist for network segmentation, *e.g.* K-means and modularity maximisation (Blondel, Guillaume, Lambiotte, and Lefebvre (2008)) are among the most popular ones.

In Figure 4, nodes with the same colour are part of the same community[1]. The results show nine different communities identified using modularity maximisation in Gephi (Bastian, Heymann, and Jacomy (2009)). The advantage of modularity maximisation is that the number of communities is defined by the algorithm. Modularity identifies nodes that are more densely connected to each other than to the rest of the network. It maximises the number of links between nodes of the same community while minimising the number of links to the rest of the network. This means that the classifications in these communities are more densely connected to each other than they are to the rest of the network. This implies that even though they are somehow related to the core of the technology, they are combined with classifications that are a bit further from the core. An example is provided in Figure 4, in which nine communities were identified; based on the structure, these were already quite easily identifiable. By reading the classifications these communities contain, we identified what these communities represent. This information is added next to the community. By reading these communities, it becomes clear that these are applications of silica but are not related to rubber. For instance, the community on the upper left (urinals) is present because the patents contain the classification of silica, and a classification that is related to non-organic compounds was also present on some of the rubber patents. The network clearly shows that these patents are out of scope.

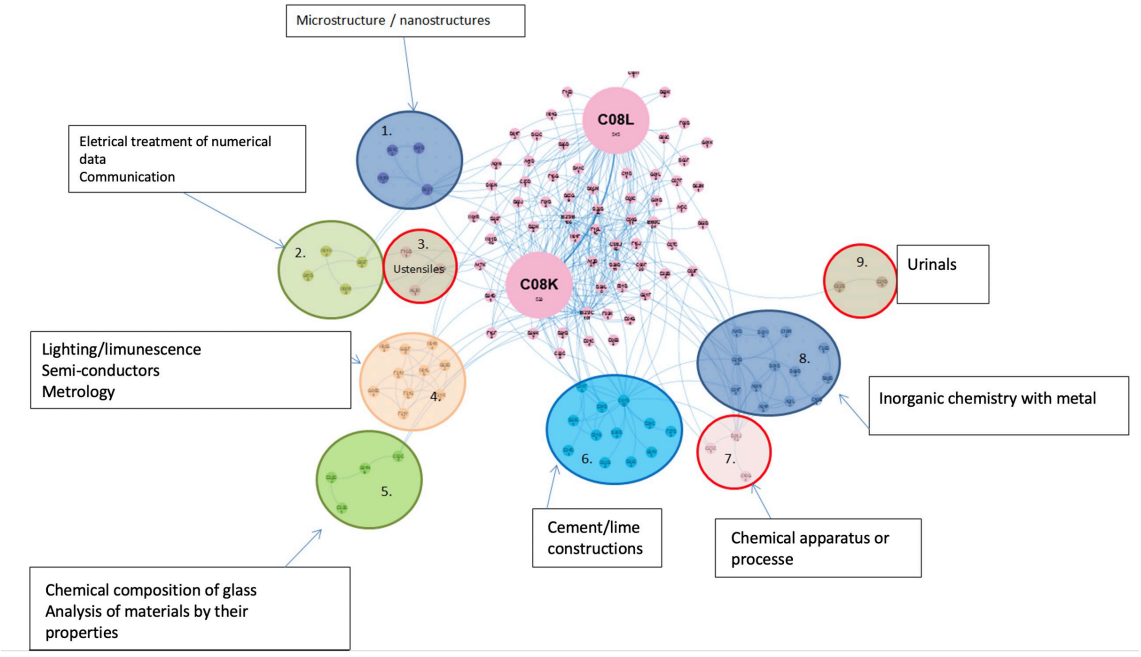

Figure 4.

However, contrary to a component, we cannot directly exclude these classifications from the query. For instance, if we were to exclude all codes from Community 1, we risk excluding a classification that would result

in other relevant patents being excluded as well. One, therefore, needs to be careful with the exclusion method used. In the present case, we would select the patents related to the core community (in the center) and keep only the patents related to the combination of those IPC codes. This means that we do not modify the query; this is done as a modification directly to the dataset.

4.3. Nodes

The third and final step of the analysis focuses on the classifications themselves. We check here for two elements: 1. How the classifications of the query are positioned. 2. We check for gatekeepers.

The network informs us on how the nodes are connected. This means that we know, if a certain classification is excluded from the query, what classifications might be affected. This step also highlights the core of our query. If a query contains classifications (as is the case for the one in our example), one should expect these classifications to be central in the network. Some classifications can play the role of gatekeeper; in other words, they connect different communities, meaning that if the community is not a false positive, this code defines the application of the technology, which is of interest since it allows for a segmentation of the query itself so one can perform an analysis on the application and the core of the technology separately.

5. Conclusion

In this paper, we show how a classification network can help researchers and analysts with the validation process of their patent query. Even though we only provide one example in this paper, we have been using this method for the purpose of producing strategic analyses of players and technological fields for multiple years now.

The method can be used on any classification as long as it is present on all patents in the set. Mostly, we have used the IPC and CPC classifications; the choice between the two mainly depends on the technological domain. In certain cases, CPC has a more precise classification system (for fuel cells, for instance).

The method reaches its limit when classification networks are very dense, which happens often in chemistry-related fields. A solution we have found is to use network reduction techniques, such as minimum spanning trees, in order to make community identification easier. This method, however, still needs work. We have not addressed the issue of false negatives in this paper. As it is, the method does not allow us to identify them. However, the proposed method can be easily implemented in recommender systems and patent query systems to improve patent information retrieval. It could even be extended to semantic networks to help with the retrieval of other types of data.

Author Contribution statement

- Jvdp + Jpr: developed the method described in the paper.
- Jvdp: wrote the manuscript text and prepared the figures.
- Jpr + Jvdp: reviewed the manuscript.

Funding Declaration

This paper has received funding from open-lab ITEMM.

Competing Interest statement

The authors have no competing interests to declare.

References

- Barabasi, A.-L. (2013). Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *371*(1987), 20120375.
- Bastian, M., Heymann, S., & Jacomy, M. (2009). *Gephi: An open source software for exploring and manipulating networks.* Retrieved from http://www.aaai.org/ocs/ index.php/ICWSM/09/paper/view/154
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.
- Coates, V., Farooque, M., Klavans, R., Lapid, K., Linstone, H. A., Pistorius, C., & Porter, A. L. (2001). On the future of technological forecasting. *Technological forecasting and social change*, *67*(1), 1–17.
- Ernst, H. (2003). Patent information for strategic technology management. *World patent information*, *25*(3), 233–242.
- Flamand, M. (2016). *Le d´eploiement de l'intelligence technologique dans le processus d'innovation des firmes: quels objectifs, enjeux et modalit´es pratiques?: Une application `a l'industrie automobileu* (Unpublished doctoral dissertation). Universit´e de Bordeaux.
- Frigant, V., & Talbot, D. (2005). Technological determinism and modularity: lessons from a comparison between aircraft and auto industries in Europe. *Industry and Innovation*, *12*(3), 337–355.
- Grabowski, H. G., & Mueller, D. C. (1972). Managerial and stockholder welfare models of firm expenditures. *The Review of Economics and Statistics*, 9–24.
- Han, F., & Magee, C. L. (2018). Testing the science/technology relationship by analysis of patent citations of scientific papers after decomposition of both science and technology. *Scientometrics*, *116*(2), 767–796.

- Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1992). *Geographic localization of knowledge spillovers as evidenced by patent citations* (Tech. Rep.). National Bureau of Economic Research.
- Khode, A., & Jambhorkar, S. (2017). A literature review on patent information retrieval techniques. *Indian Journal of Science and Technology*, *10*(36), 1–13.
- Nelson, R. R., & Winter, S. G. (1982). *An evolutionary theory of economic change*. The Belknap press of Harvard university press.
- Saint-Jean, M., Arfaoui, N., Brouillat, E., & Virapin, D. (2020). Patterns of technology knowledge in the case of ocean energy technologies. *Journal of Innovation Economics Management*, I90–33.
- Schumpeter, J. A. (1942). *Capitalism, socialism and democracy*. Routledge.
- Shalaby, W., & Zadrozny, W. (2019). Patent retrieval: a literature review. *Knowledge and Information Systems*, 1–30.
- Tijssen, R. J., Yegros-Yegros, A., & Winnink, J. J. (2016). University–industry r&d linkage metrics: validity and applicability in world university rankings. *Scientometrics*, *109*(2), 677–696.
- Trajtenberg, M. (1987). Patents, citations and innovations: tracing the links. *NBER Working Paper*(w2457).
- Trippe, A. J. (2003). Patinformatics: Tasks to tools. *World Patent Information*, *25*(3), 211–221.
- van der Pol, J., et al. (2018). *Explaining the structure of collaboration networks: from firm-level strategies to global network structure* (Tech. Rep.). Groupe de Recherche en Economie Th´eorique et Appliqu´ee (GREThA).
- van der Pol, J., & Rameshkoumar, J.-P. (2018). The co-evolution of knowledge and collaboration networks: the role of the technology life-cycle. *Scientometrics*, *114*(1), 307–323.

Declarations

Funding: This article has received funding from ITEMM-Lab

Potential competing interests: No potential competing interests to declare.