ResearchFlow: Understanding the Knowledge Flow between Academia and Industry

Angelo Salatino[0000-0002-4763-3943], Francesco Osborne[0000-0001-6557-3131], Enrico Motta[0000-0003-0015-1952]
Knowledge Media Institute, The Open University, MK7 6AA, Milton Keynes, UK
{angelo.salatino, francesco.osborne, enrico.motta}@open.ac.uk

Abstract. Understanding, monitoring, and predicting the flow of knowledge between academia and industry is of critical importance for a variety of stakeholders, including governments, funding bodies, researchers, investors, and companies. To this purpose, we introduce ResearchFlow, an approach that integrates semantic technologies and machine learning to quantifying the diachronic behaviour of research topics across academia and industry. ResearchFlow exploits the novel Academia/Industry DynAmics (AIDA) Knowledge Graph in order to characterize each topic according to the frequency in time of the related i) publications from academia, ii) publications from industry, iii) patents from academia, and iv) patents from industry. This representation is then used to produce several analytics regarding the academia/industry knowledge flow and to forecast the impact of research topics on industry. We applied ResearchFlow to a dataset of 3.5M papers and 2M patents in Computer Science and highlighted several interesting patterns. We found that 89.8% of the topics first emerge in academic publications, which typically precede industrial publications by about 5.6 years and industrial patents by about 6.6 years. However this does not mean that academia always dictates the research agenda. In fact, our analysis also shows that industrial trends tend to influence academia more than academic trends affect industry. We evaluated ResearchFlow on the task of forecasting the impact of research topics on the industrial sector and found that its granular characterization of topics improves significantly the performance with respect to alternative solutions.

Keywords: Scholarly Data, Digital Libraries, Knowledge Graph, Topic Ontology, Bibliographic Data, Topic Detection, Science of Science.

1 Introduction

Understanding, monitoring, and predicting the flow of knowledge between academia and industry is of primary importance for a variety of stakeholders, such as governments, funding bodies, researchers, investors, and companies. In particular, government and funding bodies need accurate tools to measure research impact, while companies may wish to monitor the flow of knowledge from academia to industry to ensure they stay on top of the latest scientific and innovation trends.

The complex relationship between academia and industry has been analysed from several perspectives in the literature, e.g., focusing on the characteristics of direct collaborations [1], on the influence of industrial trends on curricula [2], and the quality of the knowledge transfer [3]. However, approaches to monitoring and/or predicting the evolution of research topics typically focus either on academia [4–7] or industry [8,

9]. The few solutions that have tried to take advantage of features from both contexts have been limited to small-scale datasets, or they have focused on very specific research questions [10, 11]. Therefore, we still lack large-scale quantitative approaches to monitoring and predicting the evolution of research topics, which can integrate information from papers and patents, while also considering their provenance: academia or industry.

In this paper, we introduce ResearchFlow, a new approach for quantifying the diachronic behaviour of research topics in academia and industry. ResearchFlow builds on the Academia/Industry DynAmics (AIDA) Knowledge Graph[1] [12], a resource that we recently developed for supporting large scale analyses of academia and industry. The current version of AIDA describes 14M publications and 8M patents according to the research topics drawn from the Computer Science Ontology (CSO) [13]. Moreover, 4M publications and 5M patents are characterized according to the type of the author's affiliations (e.g., academia, industry, collaborative) and the industrial sectors (e.g., automotive, financial, energy, electronics).

ResearchFlow represents the evolution of each topic in terms of the relevant i) papers from academia, ii) papers from industry, iii) patents from academia, and iv) patents from industry. This semantic characterization takes in account the structure of the topic taxonomy described in CSO and it is used for a) producing several analytics regarding the topic evolution and the research flow between academy and industry and b) predicting the impact of research topics on the industrial sector. The resulting knowledge base, which is available at http://doi.org/10.21954/ou.rd.12805307, describes the trends of 5K topics in *Computer Science* over 2.9M papers from academia, 676K papers from industry, 2M patents from industry, and 46K patents from academia in the period 1990-2018.

The data shows that about 89.8% of the topics first appear in academic publications, 3.0% in industrial publications, and only 7.2% in patents, confirming the leading position of universities in investigating new research areas. On the average, academic publications precede industrial publications by about 5.6 years and industrial patents by about 6.6 years. However, this does not mean that academia always dictates the research agenda. In fact, if we consider only the topics for which the publication trends by academia and industry sync, after compensating for a delay, the trends from industry appear to influence academia more than academic trends influence industry. This may be due to the fact that academia tends to be quite reactive to the rise of a topic in industry (e.g., social media), which typically causes a surge of relevant academic publications in the following years. Conversely, industry appears less receptive to the emergence of topics in academia, which can be neglected for a variety of reasons – e.g., because the relevant technologies are not mature enough to support commercial products.

We evaluated ResearchFlow on the task of forecasting the impact of research topics in the industrial sector by applying several machine learning classifiers on different combinations of features. We found that the characterization of the topics produced by ResearchFlow outperforms significantly alternative solutions.

In summary, the main contributions of this paper are: i) a new approach to quantifying and forecasting the evolution of topics in academia and industry; ii) a new dataset derived from AIDA which describes the diachronic behaviour of 5K topics across 29 years (1990-2018); iii) an analysis of the patterns of knowledge flow in the

[1] Academia/Industry DynAmics Knowledge Graph - http://w3id.org/aida

field of Computer Science; and iv) a gold standard of about 39K time series that can be used for training and evaluating approaches to predicting the impact of emerging research topics on the industrial sector.

The rest of the paper is organised as follows. In Section 2, we review the literature on current approaches to studying the relationship between academia and industry, pointing out the existing gaps. In Section 3, we describe ResearchFlow and in Section 4 we provide a brief overview of the evolution of research topics in Computer Science. Section 5 reports the evaluation. Finally, in Section 6 we summarise the main conclusions and outline future directions of research.

2 Literature Review

Analysing the relationship between academia and industry allows us to understand their role within the whole knowledge economy [14]: from production, towards adoption, enrichment, and ultimately deployment as a new commercial product or service. Academia and industry typically influence each other by exchanging ideas, resources, and researchers [11]. In some cases, academia and industry engage in collaborations as an opportunity for a more productive division of tasks: academia focusing on scientific insights, and industry on commercialisation [10]. A recent book by Jack Stilgoe [15] discusses the main drivers of scientific innovation and focuses on the central role of the industry sector in pushing innovation by constantly deploying new technologies. However, it can be argued that innovation is not simply the result of the development of new technologies, but it also emerges through a more complex journey, which involves the birth of a new scientific area, the development of its theoretical framework, and the creation of innovative products that capitalise on the new knowledge [16].

So far, there has been limited investigation of this relationship. Typically, the two sectors are either analysed separately [15, 17–20] or together on a small scale [10, 11], using a limited sample of papers and patents. Most of these analyses rely on knowledge graphs describing research publications, such as Microsoft Academic Graph [21], Scopus[2], Semantic Scholar[3], Aminer [22], Core [23], OpenCitations [24], and others. Other resources, such as Dimensions[4], the United States Patent and Trademark Office corpus[5], the PatentScope corpus[6] and the European Patent Office dataset[7], offer a similar description of patents. The Semantic Web community has produces several ontologies for representing these data and the relevant research entities such as SWRC[8], BIBO[9], SPAR[10] [25], ModSci [26], and AI-KG[11] [27]. However, current knowledge graphs cannot be directly used to analyse the research dynamics of academia and industry since they lack a high quality characterization of research topics and industrial

[2] Scopus - https://www.scopus.com/
[3] Semantic Scholar - https://www.semanticscholar.org/
[4] Dimensions.ai - https://www.dimensions.ai/
[5] United States Patent and Trademark Office (USPTO) - https://www.uspto.gov/
[6] PatentScope corpus - https://patentscope.wipo.int/
[7] European Patent Office - https://data.epo.org/linked-data/
[8] SWRC - http://ontoware.org/swrc
[9] BIBO - http://bibliontology.com
[10] SPAR - http://www.sparontologies.net/
[11] AI-KG - http://w3id.org/aikg/

sectors. For this reason, we recently introduced the AIDA knowledge graph, which characterizes publications from MAG and patents from Dimensions according to the topics of CSO[12], the affiliation types of Global Research Identifier Database (GRID)[13], and the industrial sectors of the Industrial Sector Ontology (INDUSO)[14].

The relationship between academia and industry has been studied according to both qualitative and quantitative methods. A good example of the former is the work by Michaudel et al. [28] in which the authors share their personal experience on how the collaboration between industry and academia impacted their research program. Similarly, Grimpe et al. [29] performed a survey-based analysis to understand the innovation performance associated with collaborations between German manufacturers and universities. We can also find more quantitative approaches, such as Larivière et al. [30], who employed both research papers and patents to understand the primary interests of both sides in this symbiosis. Huang et al. [31] analysed 20K research papers and 8K patents in the area of *fuel cells*, in order to gain an understanding of the benefits for the two parties, which derive from industry-academic collaborations. However, all of these approaches either focus on relatively narrow areas of science or are restricted to a limited number of research questions. Other approaches focus instead on trend detection [4–6]. Typically, these methods use statistical techniques to identify, and possibly predict, the evolution of new significant areas of research. A common limitation of these techniques is that they do not take into account the types of the publications as we do. In this paper, we aim to widen the scope of this line of enquiry by developing a novel and comprehensive approach for monitoring and predicting the diffusion of research topics across academia and industry.

3 The ResearchFlow approach

The ResearchFlow approach consists of three main steps: i) generation of AIDA knowledge graph, ii) data analysis, iii) impact forecasting.

In the first phase, we generate Academia/Industry DynAmics (AIDA) Knowledge Graph, by integrating the data sources containing information about scientific articles and patents and then we enrich them by classifying documents according to i) their research topics and ii) the type of author's affiliation (academia or industry). This allows us to represent each topic according to four time series reporting the time frequency of i) papers from academia, ii) papers from industry, iii) patents from academia, and iv) patents from industry. In the second phase, we analyse the resulting time series to assess the topic trends and to identify patterns of knowledge flow. In the third phase, we use a deep learning forecaster to predict the impact of research topics.

3.1 Generation of AIDA Knowledge Graph

In order to perform a large-scale analysis of academia and industry, we need four key elements: papers, patents, research topics, and information about organizations. For this reason, we developed the AIDA knowledge graph that currently integrates 14M

[12] CSO - https://cso.kmi.open.ac.uk/
[13] Global Research Identifier Database - https://www.grid.ac/
[14] INDUSO - http://aida.kmi.open.ac.uk/downloads/induso.ttl

publications from MAG and 8M patents from Dimensions. These are described according to the topics drawn from the Computer Science Ontology (CSO) [13] and information from Global Research Identifier Database (GRID), DBpedia, and INDUSO. AIDA is generated automatically by a pipeline that is run periodically on new corpora of publications and patents. This process consists of four main steps: i) selection and integration of the relevant documents, ii) topic detection, iii) extraction of affiliation types, and iv) classification of industrial sectors.

First, we download all publications from MAG and all patents from Dimensions. MAG is a scientific knowledge base containing publication records, citations, authors, institutions, journals, conferences, and fields of study. It is one of the largest datasets of scholarly data publicly available, and, as of May 2020, it contains more than 233 million publications. Dimensions is a heterogeneous dataset containing grants, research publications, citations, clinical trials and patents. The current version includes more than 39 million patents. We then filter the resulting documents to retain only those in the field of *Computer Science*. To achieve this, we select all papers in MAG classified under "Computer Science" according to their *field of science* (FoS) [32], which is an in-house taxonomy of research areas developed by Microsoft. The patents in Dimensions are instead classified both according to the International Patent Classification (IPC) and the *fields of research* (FoR) taxonomy, which is part of the Australian and New Zealand Standard Research Classification (ANZSRC). To filter the patents in the field of Computer Science, we retain only the relevant IPC identifiers.

Since both fields of study in MAG and fields of research in Dimensions are too high level to allow a granular analysis of the knowledge flow, as a second step we annotate each paper and patent with the research topics from the Computer Science Ontology (CSO). CSO [13] is a large-scale automatically generated taxonomy of research topics in Computer Science. The current version (3.2) includes 14K research topics and 159K semantic relationships. The CSO data model is an extension of SKOS[15] and the main semantic relationships are *superTopicOf*, which is used to define the hierarchical structure of the Computer Science domain (e.g., <*artificial intelligence, superTopicOf, machine learning*>) and *relatedEquivalent,* which is used to define alternative labels for the same topic (e.g., <*ontology matching, relatedEquivalent, ontology alignment*>). We annotated publications and patents using the CSO Classifier[16] [33], an open-source Python tool for annotating documents with research topics from CSO. This is the same classifier that powers the Smart Topic Miner [34], which is the application used by Springer Nature for annotating Proceedings Book in Computer Science. The resulting set of topics was enriched by including all their super-topics in CSO. For instance, a paper tagged as *neural network* was also tagged with *machine learning* and *artificial intelligent*. This solution aims to obtain a better characterization of high-level topics that are not often directly referred in the documents.

As a third step, we classify papers and patents according to the nature of their authors' affiliations in the GRID database. GRID is a publicly available knowledge graph describing 97K organizations involved in the research. MAG and Dimensions associate the affiliations of the authors to their ID on GRID and in turn GRID associates each ID with information such as geographical location, date of establishment, alternative labels, external links, and *type of institution*, which consists of values such

[15] SKOS Simple Knowledge Organization System - http://www.w3.org/2004/02/skos
[16] CSO Classifier - https://pypi.org/project/cso-classifier/

as Education, Healthcare, Company, Archive, Nonprofit, Government, Facility, Other. We classify a document as *academia,* if all the authors have an affiliation of kind 'education' on GRID; and *industry,* if all the authors have an affiliation of kind 'company'. For the purpose of this work, we focus on these two types and ignore the collaborative efforts which constitute about 1.4% of the documents. We also do not consider the other types, which are associated with an even smaller number of documents. We plan to address both in future work.

Finally, we characterise the industrial papers and patents according to their industrial sectors. Specifically, for each industrial affiliation, we use their Wikipedia URL in GRID to query DBpedia, which is a project aiming at extracting information from Wikipedia and publish them as linked data. We exploit the predicates "About:Purpose" and "About:Industry" to retrieve the industrial sectors of each affiliation. These are then mapped to 66 main sectors described in INDUSO. Industrial sectors are not used in the current version of ResearchFlow, but they will be incorporated in the future.

AIDA is available at http://w3id.ord/aida and can be downloaded as a dump or queried via SPARQL. More details on AIDA are available in Angioni et al. [12].

3.2 Analysis

In order to focus on the main research topics, we select from AIDA only the documents associated with the most frequent *n* topics. In this paper we used *n*=5,000, resulting in 3.5M papers and 2M patents. We then associate each topic K with four time series, or signals: i) research publications from academia ($RA^K = \{RA_t^K; t \in T\}$), ii) research publications from industry ($RI^K = \{RI_t^K; t \in T\}$), iii) patents from academia ($PA^K = \{PA_t^K; t \in T\}$) and iv) patents from industry ($PI^K = \{PI_t^K; t \in T\}$), where T is the set of years considered $\{1990 \dots 2018\}$.

We perform three analyses on the resulting signals. First, we study the diachronic behaviour of topics in order to characterize their trajectory across academia and industry (Sec 3.2.1). Second, we compare each pair of signals to understand which one typically precedes the other and in which order they usually tackle a research topic (Sec 3.2.2). Finally, we assess how signals influence each other by identifying pairs of signals that are highly correlated, after compensating for a time delay (Sec 3.2.3).

3.2.1 Diachronic analysis of topics

This phase aims to quantify the evolution of a topic in previous years according to the type of documents associated with it (publications or patents) and the authors of these documents (academia or industry). For instance, we may want to detect which topics are shifting from a more academic fingerprint to a more industrial one.

As a first step we need to combine the different time series of a given topic to obtain the number of research publications (R), patents (P), documents from academia (A) and documents from industry (I) using the following formula:

$$\begin{aligned} R_t^K &= RA_t^K + RI_t^K; \quad P_t^K = PA_t^K + PI_t^K; \\ A_t^K &= RA_t^K + PA_t^K; \quad I_t^K = RI_t^K + PI_t^K; \end{aligned} \qquad t \in T$$

For example, given a topic K, its research papers time series ($R^K = \{R_t^K; t \in T\}$) is obtained by summing the number of papers from academia (RA) and industry (RI).

As second step, each point in time of each time series of each topic is normalised according to its global value for the whole Computer Science.

Therefore, given $R^{CS} = \{R_t^{CS}; t \in T\}$ the time series of research papers in Computer Science, the normalised time series of research papers R of topic K becomes:

$$R_{norm}^{K} = \left\{ \frac{R_t^K}{R_t^{CS}};\ t \in T \right\}$$

The other time series, i.e. patents (P), documents from academia (A) and documents from industry (I), are similarly obtained by combining the appropriate signals.

As a third step, we chunk our time-range in a number of time windows. For instance, if we want to observe how a particular topic changed over a period of 12 years, we may want to split it in 4 windows of 3 years. Then, for each time window *w* and for each topic K, we sum the contributions of each time series within that time window. For instance, the contribution of research papers (R_w^K) is given by:

$$R_w^K = \sum_{t=w_{init}}^{w_{end}} R_t^K$$

where w_{init} and w_{end} are the years in which the time windows respectively start and end. Similarly, we can compute the contributions of patents (P), academia (A), and industry (I).

At this stage, for a given time window, each research topic is represented by four points: total number of research publications (R_w^K), total number of patents (P_w^K), and total number of documents from academia (A_w^K) and industry (I_w^K). Then, for each topic K and for each window *w*, we define two indexes:

$$RP_w^K = \frac{R_w^K - P_w^K}{R_w^K + P_w^K}; \quad AI_w^K = \frac{A_w^K - I_w^K}{A_w^K + I_w^K}$$

The index RP allows us to observe whether in a particular time window, *w*, in proportion, a topic tends to be associated with a higher number of publications, if $RP_w^K > 0$, or patents, if $RP_w^K < 0$. The index AI instead indicates whether, in the same time window, *w*, in proportion, the topic is mostly populated by academia ($AI_w^K > 0$) or industry ($AI_w^K < 0$). In brief, for a given topic K, we now have a reduced set of time series $RP^K = \{RP_w^K; w \in W\}$ and $AI^K = \{AI_w^K; w \in W\}$, where W is the set of windows in which our initial time-frame has been divided.

In order to monitor the evolution of a topic, we can now analyse the trends of RP and AI over time. In particular, we use the least-squares approximation to determine the linear regression of both time series $f(x) = \alpha \cdot x + \beta$. Then, as trends of time series we take the slopes α^{RP} and α^{AI} of their approximated lines. If α^{RP} is positive, it means that the values in RP are growing positively over time and thus there are more papers published. On the other hand, if α^{RP} is negative it means that the number of patents is increasing in proportion. If α^{AI} is positive, it means that the topic is becoming more academic over time, whereas, if it is negative, it is becoming more industrial.

3.2.2 Analysis of Topic Emergence

In this phase, we want to asses which signal precedes another in addressing a certain topic. For instance, the topic *gamification* emerged in RA in 2008 and only five years later in RI. In the context of this analysis, we consider a topic as emergent for a certain signal when it becomes associated with at least *n* documents (*n=10* in the current implementation). Therefore, we iterate over the topics and calculate the time elapsed between the emergence of a topic for each pair of signals. Section 4.2 reports the results of this analysis on the field of Computer Science.

3.2.3 Trend Analysis

In this phase, we detect the signals that seem to influence each other by checking if they synchronize after making allowance for their mutual delay. For instance, if we consider the topic *bluetooth*, the trends of RI regularly anticipate RA, suggesting that industry is leading the research efforts for this topic. Indeed, if we align the two signals by shifting ahead RI by one year, the two signals yield a correlation coefficient $\rho = 0.975$.

In order to detect this phenomenon, we perform pairwise sliding of the time series and determine when two signals have the maximum correlation. We first normalise the time series RA^K, RI^K, PI^K and PA^K using the time series associated to the topic *Computer Science*, $RA^{CS}, RI^{CS}, PI^{CS}$ and PA^{CS}. As a second step, for each pair of time series, we compute the sliding Pearson's correlation coefficient on the overlapping part between the time series. For each couple of signals, such as RA-RI where RA is the first signal (S1) and RI the second (S2), we can define the sliding Pearson's correlation coefficient as:

$$\rho_\tau^{S1-S2} = \frac{cov(S1, S2(-\tau))}{\sigma^{S1} \cdot \sigma^{S2}}; \qquad -len(S2) + 1 \leq \tau \leq len(S1) - 1$$

where $S2(-\tau)$ is the time series of the second signal that has been shifted of $-\tau$ positions. Since $len(S2) = len(S1)$, this process produces a list of $2 \cdot len(S1) - 1$ Pearson's correlation coefficients. Having done this, we can then determine for which τ we have the highest correlation. If the maximum correlation appears for a negative value of τ, e.g. $\tau = -5$, it means that the second signal (S2:RI in the example) anticipates the first signal (S1:RA). Conversely, if τ is positive, S1 anticipates S2. However, we have observed that, within the array of correlation coefficients, there can be a number of local maxima with similar magnitude and selecting the absolute maximum may not be the appropriate solution. Therefore, to identify the value of τ that synchronises the signals we observe for which local maxima of the Pearson's correlation coefficients the two signals have also the lowest Euclidean distance.

3.3 Impact Forecasting

In this section, given the limited amount of space in this paper, we will focus specifically on predicting the impact of a research topic on the industrial landscape. Having said so, it should be emphasised that the forecaster that we have developed could indeed be used for predicting the behaviour of any of the four time series.

A good measure of the impact of a topic on industry is the number of relevant patents granted to companies. For instance, according to our data, the topic *wearable sensors* was granted only 2 patents during 2009, after which it experienced a strong acceleration, ultimately producing 135 patents in 2018. The literature proposes a wide range of approaches to patent and technology prediction through patents data, using for instance weighted association rules [9], Bayesian clustering [35], and various statistical models [36] (e.g., Bass, Gompertz, Logistic, Richards). In the last few years, we saw also the emergence of several approaches based on Neural Networks [8, 37], which often yield the most competitive results. However, most of these tools focus only on patents, and do not integrate research publication data, nor can they distinguish patents and publications produced by academia or industry.

The ResearchFlow approach can naturally support all these solutions since it produces a large quantity of granular data that can be used to train and test machine

learning classifiers. Furthermore, we hypothesize that an input which integrates all the information about publications and patents should offer a richer set of features and would be more robust in situations in which patents data are scarce, ultimately yielding a better performance in comparison to approaches which rely solely on patent data.

In order to train a forecaster, we created a gold standard, in which for each topic in CSO, we selected all the time-frames of five years in which the topic had not yet emerged (less than 10 patents). We then labelled each of these samples as *True* if the topic produced more than 50 industrial patents (PI) in the following 10 years and *False* otherwise. The resulting dataset includes 9,776 labelled samples, each composed of four time series (RA, RI, PA, PI). We then implemented a neural network forecaster which uses one Long short-term memory (LSTM) hidden layer of 128 units and one output layer computing the softmax function. We use binary cross-entropy as loss function and train the model over 50 epochs. Section 5 reports the evaluation of this architecture versus alternative approaches.

4 Results from the analysis of Computer Science

We used ResearchFlow to quantify the trends of 5K topics in *Computer Science* over 2.9M research papers from academia (RA), 676K research papers from industry (RI), 2M patents from industry (PI), and 46K patents from academia (PA) in the period 1990-2018. Because of space restrictions, we will focus the discussion only on the main insights that emerged from our experiments.

4.1 Diachronic analysis

Fig. 1 shows the distribution of all topics in a 2-dimensional diagram with AI on the horizontal axis and RP on the vertical axis (computed as described in Section 3.2.1). Interestingly, most topics are tightly distributed around the bisector.

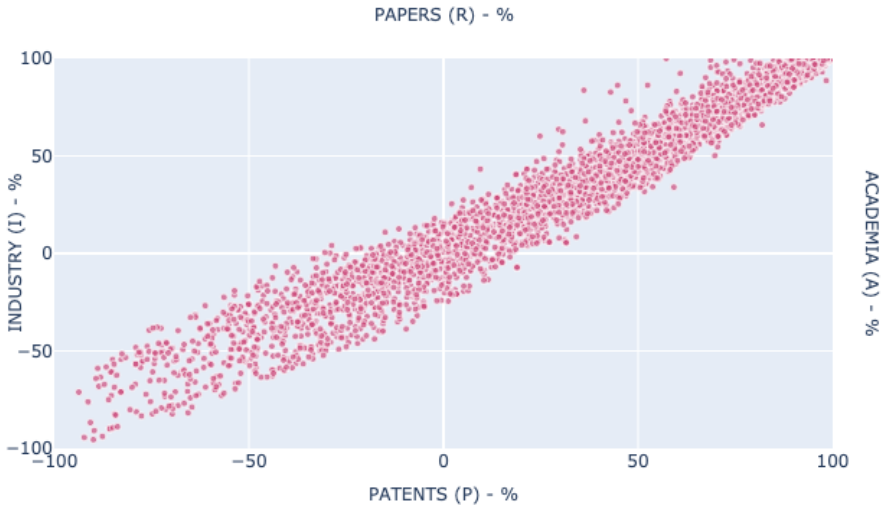

Fig. 1. Top 5,000 topics in Computer Science according to their RP and AI indexes.

The topics which attract most interest from academia mainly produce research papers (top-right quadrant). Conversely, the topics which are more interesting for industry tend to generate prevalently patents (bottom left quadrant). This distribution follows a classic pattern, consistent with the analysis of Larivière et al. [30], which suggest that academia is mostly interested on the dissemination of knowledge through scientific articles, while companies focus more on preserving their intellectual property by producing patents.

In the top-right quadrant we find research topics, such as *e-learning systems*, *scholarly communication, smart environment, community detection, decision tree algorithms*, which are mostly populated by academics. In the bottom-left quadrant we tend to find more applied areas, such as *optoelectronic devices*, *high power lasers*, *network interface*, *flip-flop*, *optical signals, magnetic disk storage*.

We applied the diachronic topic analysis described in Section 3.2.1 to highlight the topics that experienced the most dramatic shift in this space. We focused on the last 12 years (2007-18), using 4 windows of 3 years each. Table 1, Table 2, Table 3, and Table 4 report the top 5 topics that have respectively the strongest trends towards publications ($\alpha^{RP} > 0$), patents ($\alpha^{RP} < 0$), academia ($\alpha^{AI} > 0$), and industry ($\alpha^{AI} < 0$). We also report the values of the two indexes (RP and AI) in the first (2007-2009, RP_1 and AI_1) and the last (2016-2018, RP_4 and AI_4) time windows. Although it is not possible without additional analysis to come to definitive conclusions, these tables provide valuable information by highlighting areas of relative high/low activity.

Overall, the top five entries which had a strong increment in the direction of academia and publications (Table 1 and 3) can be categorized in three macro areas: energy production (e.g., *smart grid, energy harvesting*), technologies for telecommunication (e.g., *internet of things, slot antennas*), and data security (e.g., *encrypted data*). Conversely, the main entries for industry and patents (Table 2 and 4) focus prevalently on technologies for telecommunication (e.g., *overlay networks, long term evolution, coding mode*), user interfaces (e.g., *hand gesture, wearable computing*), and image processing (e.g., *video encoder, 3d video*).

Table 1. Topics with strongest trends towards publications.

Topic	α^{RP}	RP_1	RP_4
smart grid	27.2	-21.1	65.1
internet of things	26.6	-8.5	76.8
energy harvesting	23.3	-58.1	13.8
matrix factorization	22.2	6.8	72.1
slot antennas	22.1	-52.5	18.7

Table 2. Topics with strongest trends towards patents.

Topic	α^{RP}	RP_1	RP_4
long term evolution (lte)	-31.0	89.0	-0.9
mode decision (coding)	-27.7	46	-36.2
3d video	-26.9	72.5	-4.1
overlay networks	-25.2	81.5	6.8
hand gesture	-23.1	59.1	-6.5

Table 3. Topics with strongest trends towards academia.

Topic	α^{AI}	AI_1	AI_4
smart grid	26.9	-14.2	68.5
internet of things	25.2	-6.0	68.9
encrypted data	24.9	-62.4	9.88
distribution systems	23.4	-17.9	52.9
energy harvesting	22.1	-44.9	22.7

Table 4. Topics with strongest trends towards industry.

Topic	α^{AI}	AI_1	AI_4
overlay networks	-21.8	72.8	7.5
mode decision (coding)	-21.5	30.4	-34.5
long term evolution (lte)	-19.2	52.4	-3.2
wearable computing	-18.6	72.8	16.08
video encoder	-17.1	-14.1	-66.2

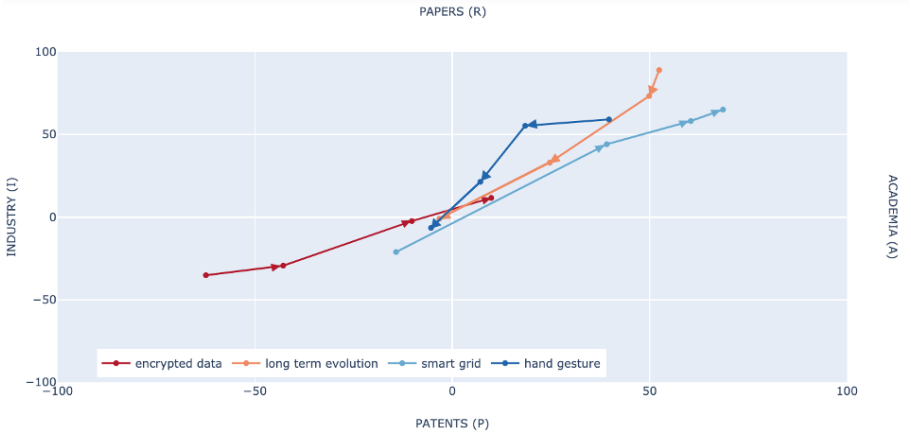

Fig. 2. Trajectories of *encrypted data*, *long term evolution*, *smart grid* and *hand gesture*.

Fig. 2 shows as example the trajectories of four topics that exhibited a dramatic shift in the period 2007-2018: *encrypted data*, *long term evolution* (a standard for broadband wireless technology), *smart grid*, and *hand gesture*. *Encrypted data* (red line) was in the left-bottom area, which characterizes prominently industrial topics, counting 178 documents from academia (A) and 560 from industry (I) in the first windows (2007-09), before being increasingly adopted by academia and moving up to the top-right area, counting A = 894 and I = 453 in the last window (2016-18). *Smart grid* (light blue line) followed a similar trajectory. On the other hand, *long term evolution* (orange line) and *hand gesture* (dark blue line) followed the opposite trajectory. Specifically, in the first window *hand gesture* was primarily an academic topic, counting A = 1,107 and I = 348; it then became more and more industrial over the years, increasing the number of documents from academia to 2,218, and from industry to 2,133. Similarly, *long term evolution* was initially in the top-right quadrant finding more industrial application over time as it became a well adopted standard.

4.2 Analysis of Topic Emergence

In this section we report the results of the analysis described in Sec. 3.2.2 on the 3,484 topics that according to their four associated signals emerged after 1990, which is the first year of our dataset.

We found that 89.9% of the topics first emerge in academic publications, 3.0% in industrial publications, 7.2% in industrial patents, and none in academic patents. On average, publications from academia (RA) precede publications from industry (RI, see Fig. 3) by 5.6±5.6 years, and in turn RI precedes patents from industry (PI, see Fig. 4) by 1.0±5.8 years. RA also precedes by 6.7±7.4 years patents from industry (PI, see Fig. 5). However, just considering the average would be misleading in this case. Indeed, as depicted by Fig. 3, in 15.7% of cases the topics emerged in RI only one year later than RA, and in the 11.7% two years later.

For the sake of space we do not show the distributions involving PA, that counts only 1,897 emerging topics. An analysis of this set showed that topics in PA appear on average 20.4±7.0 years after they emerge in RA, 14.8±7.2 after RI, and 13.7±7.3 after

PI. In conclusion, these results confirm that the academia is usually the first to investigate a topic and suggest that industrial publications are conducive to patents.

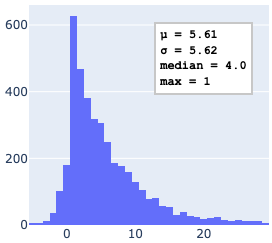

Fig. 3. S1:RA - S2:RI

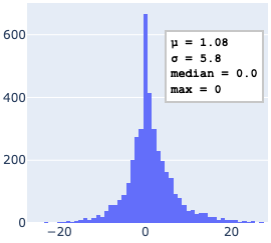

Fig. 4. S1:RI - S2:PI

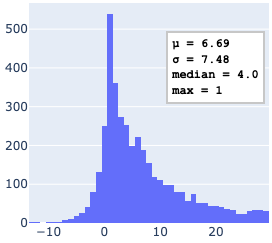

Fig. 5. S1:RA - S2:PI.

4.3 Trend Analysis

We performed the analysis described in section 3.2.3 on all the topics and determined the time delay (τ) between each pair of time series S1 and S2. The following figures show the distributions of the delay for the six pairwise comparisons between the four time series. The x-axis represents the time lag τ, while the y-axis represents the number of topics in which the maximum Pearson's correlation coefficient was found in τ. We included only maxima in which $\rho \geq 0.7$, which is traditionally considered a strong direct correlation. We remind the reader that, as per our convention, the signal S2 is sliding over S1, and a maximum correlation in a negative τ means that S2 anticipates S1. Conversely, a positive τ means that S1 anticipates S2.

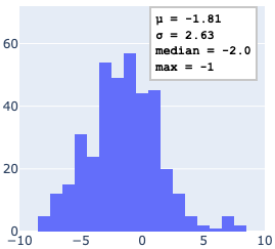

Fig. 6. S1:RA - S2:RI

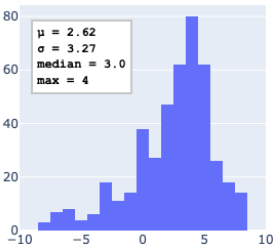

Fig. 7. S1:RI - S2:PI

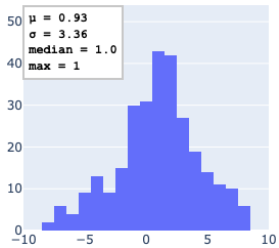

Fig. 8. S1:RA -S2:PI

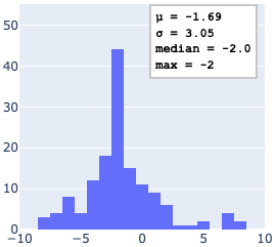

Fig. 9. S1:PA - S2:RA

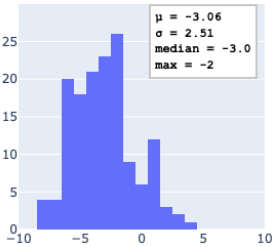

Fig. 10. S1:PA - S2:RI

Fig. 11. S1:PA - S2:PI

Fig. 6 shows that when we consider only the 327 topics for which RA and RI sync after compensating for a delay, the trends of RI tend to anticipate the ones of RA by almost 1.8 years on the average. In other words, an increasing interest of the industrial

sector appears to often trigger a reaction in academia, while the opposite case is less frequent. A more in depth analysis on the involved topics seems to suggest that this is due to the fact that academia often reacts to the emergence of a topic in industry (e.g., social media, mobile devices, internet of things) by further investigating it. Conversely, industry tends to be less receptive and in some cases to ignore or react slowly to the emergence of topics in academia. This asymmetry is an intriguing phenomenon that we intend to further investigate in future works.

Another interesting dynamics is that the trends of industrial patents (PI) are anticipated by the trends of publications from industry (RI) with a delay of about 2.6 years (Fig. 7) and by academic publications (RA) by 1 year (Fig. 8). This suggests that both could be good predictors for patents. Finally, on average patents from academia (PA) tend to sync with publications from academia with a delay of almost 1.7 years (Fig. 9), industrial publications by 3.0 years (Fig. 10), and industrial patents by 0.4 year (Fig. 11).

5 Evaluation

In order to verify the hypothesis that a forecaster which integrates all the signals produced by ResearchFlow will yield better performance than the systems [8, 35–37] that utilize only the number of publications or patents, we evaluated several models on the task of predicting if an emergent research topic will have a significant impact on the industrial sector, producing more than 50 patents in the following 10 years. We thus trained five machine learning classifiers on the gold standard introduced in Section 3.3: Logistic Regression (LR), Random Forest (RF), AdaBoost (AB), Convoluted Neural Network (CNN), and Long Short-term Memory Neural Network (LSTM). We ran each of them on research papers (R), patents (P), and the 15 possible combinations of the four time series in order to assess which set of features would yield the best results. We performed a 10-fold cross-validation of the data and measured the performance of the classifiers by computing the average precision (P), recall (R), and F1 (F). The dataset, the results of experiments, the parameter and implementation details, and the best models are available at http://doi.org/10.21954/ou.rd.12805307.

Table 5 shows the results of the evaluation. We report all combinations in order to assess the contributions of the different time series. LSTM outperforms all the other solutions, yielding the highest F1 for 12 of the 17 feature combinations and the highest average F1 (73.7%). CNN (72.8%) and AB (72.3%) also produce competitive results. For the sake of space, here we will focus on the performance of the LSTM models.

As hypothesized, using the full set of features produced by ResearchFlow (RA-RI-PA-PI) significantly ($p<0.0001$) outperforms (F1: 84.6%) the solution which uses only the number of patents by companies (74.8%). Splitting each of the two main time series (publications and patents) in its components (academia and industry) also increases performance: RA-RI (80.7%) significantly ($p<0.0001$) outperforms R (68.2%) while PA-PI (75.2%) is marginally better than P (74.8%). This confirms that the more granular representation of the document origin can increase the forecaster performance.

When considering the models produced with only one of the time series, we find that the number of publications from industry (RI) is a significant ($p=0.004$) better indicator than PI, yielding a F1 of 76.9%, followed by RA, and PA. If we zoom on the models trained on two time series, the best results are obtained by the combinations RI-PI (81.4%), when considering three, RA-RI-PI yields the best performance (84.7%).

In conclusion, this evaluation substantiates the hypothesis that considering the four time series separately is conducive to higher quality predictions and suggests that RI and RA are good indicators for PI.

Table 5. Performance of the five classifiers on 17 combinations of time series. In bold the best F1 (F) for each combination.

	LR			RF			AB			CNN			LSTM		
	P%	R%	F%	P%	R%	F%	P%	R%	F%	P%	R%	F%	P%	R%	F%
RA	70.8	45.2	55.2	63.3	55.8	59.2	66.0	58.4	61.9	64.1	66.3	**65.0**	65.2	64.2	64.6
RI	83.5	67.1	74.4	78.9	69.8	74.0	80.0	73.1	76.4	79.2	75.1	**77.0**	79.1	74.8	76.9
PA	58.3	15.3	24.2	60.4	15.4	24.5	59.3	16.0	**25.2**	60.5	15.7	24.9	60.8	15.6	24.8
PI	76.5	69.0	72.5	73.9	68.4	71.0	75.6	71.8	73.6	73.7	76.6	75.0	74.1	76.6	**75.2**
R	73.7	48.8	58.7	65.5	59.7	62.5	68.6	63.1	65.6	67.6	69.2	**68.3**	67.2	69.4	68.2
P	76.5	68.6	72.3	72.8	67.6	70.0	74.4	71.6	73.0	73.2	76.1	74.6	73.1	76.6	**74.8**
RA-RI	85.7	70.9	77.6	80.5	76.0	78.2	82.6	76.6	79.5	78.9	75.1	76.8	82.2	79.3	**80.7**
RA-PA	70.3	47.0	56.3	63.1	55.5	59.0	66.5	59.3	62.6	64.5	65.1	64.5	65.4	64.2	**64.6**
RA-PI	79.6	73.7	76.5	77.2	74.3	75.7	79.1	76.5	77.7	75.2	76.3	75.7	77.4	81.9	**79.5**
RI-PA	83.3	67.0	74.3	77.9	70.8	74.1	79.6	73.0	76.1	78.6	75.6	77.0	79.1	75.2	**77.1**
RI-PI	83.4	77.3	80.2	81.0	77.3	79.1	82.7	78.6	80.6	82.0	78.6	80.2	81.7	81.2	**81.4**
PA-PI	76.7	68.6	72.4	74.2	69.0	71.5	75.9	71.5	73.6	71.1	70.8	70.9	73.8	76.7	**75.2**
RA-RI-PA	85.2	71.4	77.7	80.8	75.4	78.0	82.5	77.0	79.6	82.6	78.1	**80.3**	82.6	78.2	**80.3**
RA-RI-PI	85.4	79.8	82.5	84.5	80.5	82.4	84.6	81.2	82.9	83.8	84.7	84.2	84.1	85.4	**84.7**
RA-PA-PI	79.6	73.9	76.6	77.5	74.4	75.9	79.2	76.5	77.8	78.9	78.6	78.6	77.4	81.4	**79.2**
RI-PA-PI	83.6	77.5	80.4	81.1	78.0	79.5	82.7	78.6	80.6	82.2	80.9	**81.5**	81.1	81.0	81.1
RA-RI-PA-PI	85.4	79.8	82.5	83.8	80.0	81.8	84.6	81.2	82.9	84.7	81.3	82.9	83.2	86.1	**84.6**
Average	78.7	64.8	70.2	75.1	67.5	70.4	76.7	69.6	72.3	75.4	72.0	72.8	75.7	73.4	**73.7**

6 Conclusions and Future Work

In this paper, we introduced ResearchFlow, an approach to analysing and forecasting the knowledge flows between academia and industry. We applied ResearchFlow on a dataset of publications and patents in Computer Science, and produced a knowledge base that described the behaviour of topics across academia and industry. Our analysis indicates that academia is the first to investigate most of these topics; on the average, academic publications precede industrial publications by about 5.6 years and industrial patents by about 6.6 years. However, industrial trends actually appears to influence academia more often than academic trends affect industry, suggesting that in several cases it is industry that dictates the research direction. Finally, we showed that quantifying research topics according to the four time series described in this work can significantly increase the performance of a forecaster.

We are now working on a more comprehensive analysis of Computer Science which will include the full range of analytics that we can produce with ResearchFlow and a more detailed discussion. In particular, we intend to investigate further the specific mechanisms that allow industry to influence academia and the other way round. We also intend to analyse documents with mixed affiliations and extend this analysis to other kinds of organisations, such as healthcare, government, and non-profit.

References

1. Ankrah, S., AL-Tabbaa, O.: Universities-industry collaboration: A systematic review. Scand. J. Manag. 31, 387–408 (2015). https://doi.org/10.1016/j.scaman.2015.02.003.
2. Weinstein, L., Kellar, G., Hall, D.: Comparing Topic Importance Perceptions of Industry and Business School Faculty: Is the Tail Wagging the Dog? Acad. Educ. Leadersh. J. 20, 62 (2016).
3. Ankrah, S.N., Burgess, T.F., Grimshaw, P., Shaw, N.E.: Asking both university and industry actors about their engagement in knowledge transfer: What single-group studies of motives omit. Technovation. 33, 50–65 (2013).
4. Ohniwa, R.L., Hibino, A., Takeyasu, K.: Trends in research foci in life science fields over the last 30 years monitored by emerging topics. Scientometrics. 85, 111–27 (2010).
5. Salatino, A.A., Osborne, F., Motta, E.: AUGUR: Forecasting the Emergence of New Research Topics. In: Joint Conference on Digital Libraries 2018, Fort Worth, Texas. pp. 1–10 (2018).
6. Bolelli, L., Ertekin, Ş., Giles, C.L.: Topic and trend detection in text collections using latent dirichlet allocation. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 5478 LNCS, 776–780 (2009).
7. Salatino, A.A., Osborne, F., Motta, E.: How are topics born? Understanding the research dynamics preceding the emergence of new areas. PeerJ Comput. Sci. 3, e119 (2017). https://doi.org/10.7717/peerj-cs.119.
8. Zang, X., Niu, Y.: The forecast model of patents granted in colleges based on genetic neural network. In: 2011 International Conference on Electrical and Control Engineering, ICECE 2011 - Proceedings. pp. 5090–5093 (2011).
9. Altuntas, S., Dereli, T., Kusiak, A.: Analysis of patent documents with weighted association rules. Technol. Forecast. Soc. Change. 92, 249–262 (2015).
10. Bikard, M., Vakili, K., Teodoridis, F.: When Collaboration Bridges Institutions: The Impact of University–Industry Collaboration on Academic Productivity. Organ. Sci. 30, 426–445 (2019). https://doi.org/10.1287/orsc.2018.1235.
11. Anderson, M.S.: The complex relations between the academy and industry: Views from the literature. J. Higher Educ. 72, 226–246 (2001). https://doi.org/10.2307/2649323.
12. Angioni, S., Salatino, A., Osborne, F., Recupero, D.R., Motta, E.: Integrating Knowledge Graphs for Analysing Academia and Industry Dynamics. In: ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium. Springer International Publishing (2020).
13. Salatino, A.A., Thanapalasingam, T., Mannocci, A., Osborne, F., Motta, E.: The Computer Science Ontology : A Large-Scale Taxonomy of Research Areas. In: The Semantic Web -- ISWC 2018. Springer (2018).
14. Powell, W.W., Snellman, K.: The Knowledge Economy. Annu. Rev. Sociol. 30, 199–220 (2004). https://doi.org/10.1146/annurev.soc.29.010202.100037.
15. Stilgoe, J.: Who's driving innovation? new technologies and the collaborative state. Palgrave Macmillan (2020).
16. Kuhn, T.S.: The structure of scientific revolutions. University of Chicago Press (2012).
17. Becher, T., Trowler, P.: Academic tribes and territories : intellectual enquiry and the culture of disciplines. Open University Press (2001).
18. Krumov, L., Fretter, C., Müller-Hannemann, M., Weihe, K., Hütt, M.-T.: Motifs in co-authorship networks and their relation to the impact of scientific publications. Eur. Phys. J. B 2011 844. 84, 535–540 (2011). https://doi.org/10.1140/EPJB/E2011-10746-5.
19. Varlamis, I., Tsatsaronis, G.: Visualizing bibliographic databases as graphs and mining potential research synergies. In: Proceedings - 2011 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011. pp. 53–60 (2011). https://doi.org/10.1109/ASONAM.2011.52.
20. Frank, M.R., Wang, D., Cebrian, M., Rahwan, I.: The evolution of citation graphs in

artificial intelligence research, https://www.nature.com/articles/s42256-019-0024-5, (2019). https://doi.org/10.1038/s42256-019-0024-5.
21. Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., Kanakia, A.: Microsoft Academic Graph: When experts are not enough. Quant. Sci. Stud. 1, 396–413 (2020).
22. Zhang, Y., Zhang, F., Yao, P., Tang, J.: Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop. KDD. 18, 10.
23. Knoth, P., Zdrahal, Z.: CORE: Three access levels to underpin open access. D-Lib Mag. 18, (2012). https://doi.org/10.1045/november2012-knoth.
24. Peroni, S., Shotton, D.: OpenCitations, an infrastructure organization for open scholarship. Quant. Sci. Stud. 1, 428–444 (2020). https://doi.org/10.1162/qss_a_00023.
25. Peroni, S., Dutton, A., Gray, T., Shotton, D.: Setting our bibliographic references free: towards open citation data. J. Doc. 71, 253–277 (2015).
26. Fathalla, S., Auer, S., Lange, C.: Towards the Semantic Formalization of Science. (2020). https://doi.org/10.1145/3341105.3374132.
27. Dessì, D., Osborne, F., Recupero, D.R., Buscaldi, D., Motta, E., Sack, H.: AI-KG: an Automatically Generated Knowledge Graph of Artificial Intelligence. In: The Semantic Web -- ISWC 2020. Springer Verlag (2020).
28. Michaudel, Q., Ishihara, Y., Baran, P.S.: Academia-Industry Symbiosis in Organic Chemistry. Acc. Chem. Res. 48, 712–721 (2015). https://doi.org/10.1021/ar500424a.
29. Grimpe, C., Hussinger, K.: Formal and Informal Knowledge and Technology Transfer from Academia to Industry: Complementarity Effects and Innovation Performance. Ind. Innov. 20, 683–700 (2013). https://doi.org/10.1080/13662716.2013.856620.
30. Larivière, V., Macaluso, B., Mongeon, P., Siler, K., Sugimoto, C.R.: Vanishing industries and the rising monopoly of universities in published research. (2018).
31. Huang, M.H., Yang, H.W., Chen, D.Z.: Industry–academia collaboration in fuel cells: a perspective from paper and patent analysis. Scientometrics. 105, 1301–1318 (2015).
32. Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. (Paul), Wang, K.: An Overview of Microsoft Academic Service (MAS) and Applications. In: Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion. pp. 243–246. ACM Press, New York, New York, USA (2015).
33. Salatino, A.A., Osborne, F., Thanapalasingam, T., Motta, E.: The CSO Classifier: Ontology-Driven Detection of Research Topics in Scholarly Articles. In: TPDL 2019: 23rd International Conference on Theory and Practice of Digital Libraries. Springer.
34. Salatino, A.A., Osborne, F., Birukou, A., Motta, E.: Improving Editorial Workflow and Metadata Quality at Springer Nature. In: The Semantic Web – ISWC 2019. Springer Verlag (2019).
35. Choi, S., Jun, S.: Vacant technology forecasting using new Bayesian patent clustering. Technol. Anal. Strateg. Manag. 26, 241–251 (2014).
36. Marinakis, Y.D.: Forecasting technology diffusion with the Richards model. Technol. Forecast. Soc. Change. 79, 172–179 (2012).
37. Ramadhan, M.H., Malik, V.I., Sjafrizal, T.: Artificial neural network approach for technology life cycle construction on patent data. In: 2018 5th International Conference on Industrial Engineering and Applications, ICIEA 2018. pp. 499–503 (2018).