Automatically identifying emerging technologies using open source data

1 Introduction

Cyberrisk management largely reduces to a race for information between defenders and attackers of ICT systems. Under the threat of costly attacks, defenders can gain an advantage with the early identification and adoption of new cyberdefense technologies. Cybersecurity spans a wide range of fields, including computer science, data science, telecommunications and embedded devices, and thus the need for comprehensive analytics on emerging technologies. Such knowledge is crucial for companies, research institutions and large governments agencies seeking to shore up their defense systems and understand the attack surfaces of the future. A technology roadmap thus helps these stakeholders allocate resources in research and development and also allows them to purchase the most adequate systems from third-parties. This strong need for insights on emerging technologies is currently answered by wide media coverage on the topic and has driven the business of companies like Gartner and Forrester providing technology forecasting research.

Despite the common and wide-spread use of the term *emerging technologies* there is no strict definition that would allow for an easy distinction between emerging, established and declining technologies [27]. The absence of such a definition makes it difficult to develop scientifically sound identification methods. Gartner's famous Hype Cycle for Emerging Technologies appears very intuitive but cannot really be used as an underlying model. In the literature, Gartner's concept is often criticized as unscientific, inconsistent, generic, and subjective [30, 4]. Other research organizations like Forrester, IHS Markit, etc. also produce yearly reports of emerging technologies. However, the methodology used by these organizations for identifying emerging technologies remains unknown.

Research in the area of identifying emerging technologies is limited to qualitative methods, expert systems and survey based methods. For quantitative methods, scholars have used open data sets and S-curve models for identifying technology emergence [1, 16, 24, 26]. S-Curve models are based on the concept of logistic or Gompertz growth that eventually leads to saturation. The clear advantage of such models is their sound mathematical foundation. After fitting the data to the model, the exact growth formula allows to assess the current maturity state of a technology and its future development. However, the majority of these studies [16, 24, 26] focuses on a specific set of technologies chosen in advance. As a consequence, models are specific to particular technologies and a general method for detecting emerging technologies is difficult to devise.

In this paper, we first seek to test whether a purely quantitative approach is feasible in order to identify emerging technologies on a large scale using open data. This comes as an alternative to the qualitative models mentioned above. Secondly, we will examine whether manually-selected indicators perform as well as Machine Learning processes to find relationships and technology clusters in the data. The precision and accuracy scores of both methods will help us compare their performances.

To the best of our knowledge, none of the prior research in this field has combined all these data sources into a single method. Also, unlike prior research, we have not selected any particular set of technologies in advance. Thus, we develop a scientifically-sound, scalable and automatic technology predictors that allow to (1) accelerate the early identification of technologies relevant for cyberdefense, (2) increase the adoption time of a given identified technology and (3) reduce the technological information asymmetry between the attacker and the defender.

The remainder of this paper is organized as follows. Section 2 provides a survey on existing research on identifying emerging technologies. In section 3 we give a detailed description of our data acquisition process. Section 4 describes the first applied methodology, which uses manually selected criteria to evaluate technologies. We then present the machine learning classifier in section 5. We discuss and compare the results of both predictors in Section 6. Section 7 concludes the paper and provides an outlook on future work.

2 Literature survey

While the meaning of the term 'emerging' seems intuitive, it is difficult to recognize a commonly accepted definition. This is reflected by the large number of publications that attempt to do so. The definitions proposed in the literature overlap but use different characteristics. Some authors [25, 21, 7, 15, 12] emphasize the technology's potential impact on the economy or the society including both evolutionary change as well as disruptive innovations. Other scholars [3] regard the aspect of uncertainty about the future evolution of a technology as more important. Some other authors combine both aspects potential and uncertainty [8, 29], and yet others [28] underline novelty and growth.

The variety of chosen characteristics for emerging technologies has led to a number of different, mostly scientometric approaches for measurement [11] which lack a clear definition of the underlying concept of emergence. Rotolo, Hicks and Martin [27] conveyed a thorough analysis of existing research on the definition of emerging technologies and aggregated comparable approaches. They provide a thorough analysis of different approaches for defining emerging technologies and derives five main characteristics, a subset of which commonly appears across the studied research. These are (i) radical novelty, (ii) relatively fast growth, (iii) coherence, (iv) prominent impact, and (v) uncertainty. We use this definition as a starting point for our investigation and create from these five concepts relevant predictors for the first methodology.

Publicly available data is often used to predict emerging technologies. Commonly exploited data set are patents such as United States Patent and Trademark Office (USPTO), Global Patent Index (GPI), and Thompson Innovation.

Many publications propose to use bibliometric methods to extract data and identify emerging technologies, and then deploy growth models for technology prediction. In [9] they applied bibliometric methods, US patent analysis and S-curves for forecasting fuel cells, food safety, and optical storage using a technology-specific set of methods. Similarly, [26] used expert interviews to fit data acquired by text-mining patents into growth curve models for predicting hybrid cars and fuel cells. Text-mining on patents and fitting to S-curves was also proposed in [18], and [2] found correlation between patent and publication data extracted by scientometric methods for 20 technologies and deployed S-curves for forecasting. S-Curve models for predicting emerging technologies were also proposed by [16, 24].

Recently artificial intelligence has (re-)gained much attention and consequently machine learning has been used to model and predict emerging technologies. [19] used supervised learning on citation graphs from USPTO data to automatically label and forecast emerging technologies with high precision in a given year. Similarly, [32] applied supervised deep learning on world-wide patent data. The training sets were labelled manually based on Gartner's Hype Cycle. [20] extracted 21 indicators from the USPTO data and using neural networks achieved impressive predictive power on a subset of 35,256 patents. All this prior research will inspire our own machine learning classifier presented in Section 5.

Other companies and institutions, besides Gartner, publish lists of emerging technologies on a regular basis. The exact selection criteria that resulted in the publication of these lists remain unclear. Despite differences both in terms of the frequency of publication and the length of the emerging technology lists, an important overlap can be noticed. We use the union set of the lists published by the companies listed below as a baseline for performance comparison.

Gartner[1] is a global market analysis company which provides insights for a wide range of industries. Widely known are their branded visualizations for market research reports such as the Hype Cycle and Magic Quadrants.

Forrester Research[2] is an American market research company that provides advice on existing and potential impact of technology. Forrester publishes lists containing 15 emerging technologies for the next 5 years with the last part in 2016 for the time until 2021.

IHS Markit is a global information provider with headquarters in London that was formed in 2016 from the merger of IHS Ltd. and Markit Ltd. Their track record for identifying emerging technologies is rather small. In 2017 they published 7 transformative technologies[3] followed by 8 technologies[4] for 2018. The two lists have overlaps and there is a remarkable co-occurrence of AI and cloud technology for both years.

[1] https://www.gartner.com/smarterwithgartner/top-trends-from-gartner-hype-cycle-for-digital-government-technology-2018

[2] https://go.forrester.com/blogs/top-emerging-technologies-2018/

[3] https://ihsmarkit.com/research-analysis/7-in-2017-top-tech-trends-this-year.html

[4] https://ihsmarkit.com/Info/1217/top-transformative-technology-trends-2018.html

World Economic Forum (WEF): Since 2011 WEF's Global Agenda Council on Emerging Technologies, a volunteer network of experts, has been publishing a yearly list of the top-10 emerging technologies [5].

3 Data sources

This section presents the data sources used in our study to identify emerging technologies. All data sets are open to the public and available free of charge.

Patents and publications provide essential means to capture growth and novelty of a technology [1, 23, 14]. Researchers have used them in the past in various ways to capture the growth of a technology based on its time series [2, 9]. In our study we have used the Patentsview data set for patents and the arXiv data set for publications. Emerging technologies have a high and quick public acceptance rate [21, 10, 25] due to the high impact on consumers. Sources like Google search represent web search behaviour and public interest in a particular topic [6, 31, 5, 17], thereby generating statistics for Google trends. These searches often lead to the Wikipedia page for detailed information about the topic. Therefore, it is conceivable that Google trends provide knowledge about initial queries which users are looking for, whereas Wikipedia page views capture deeper interest of the public in a specific topic or field of study [17]. Hence, Wikipedia page views should indicate the impact of a technology on the public. Moreover, Wikidata enables for the linkage between the Wikipedia pages.

Patentsview provides the information of patents from the USPTO for all applications since 1976. The bulk data set is available for download on their website. We fetched the data set in May 2018 which then contained 6,647,699 patents. Each patent has a unique identifier and a grant date, which we use to create time series for technologies.

ArXiv is a repository for scientific papers in many science and engineering areas. ArXiv provides an API for downloading all publications over the OAI (Open Archives Initiative) protocol. We fetched the bibliographic information of 1,425,558 papers until September 2018 including publication date and abstract.

Wikipedia Page Views provides open access to the page view statistics through an API. We fetched the Wikipedia monthly views of the project *en.wikipedia* for 50,954 articles identified as technology by the classifier (see section 4.1). Furthermore, we computed the relative page view of an article with respect to the total number of page views to compensate for the increased popularity of Wikipedia over the years.

Wikidata is a free and open knowledge base that can be read and edited by both humans and machines. Wikidata acts as central storage for the structured data of its Wikimedia sister projects.We used the Wikidata properties[6] for linking the retrieved Wikipedia technology articles in the evaluation phase.

[5] https://www.weforum.org/agenda/2018/09/top-10-emerging-technologies-of-2018

[6] https://www.wikidata.org/wiki/Wikidata:List_of_properties

4 Methodology

In this section we describe the methodology used for technology classification, data extraction, technology time series creation as well as the emergence scoring.

Figure **??** gives an overview of the methodology used for identifying and scoring technologies for emergence. We quantified 4 out of 5 criteria of an emerging technology as stated in [27]: novelty, growth, impact, and coherence. We did not provide a metric for uncertainty which is not deterministic by definition.

We define as a technology any Wikipedia article classified as such by our algorithm described in section 4.1. Using DBPedia, containing a multi-domain ontology derived from Wikipedia and Spotlight [22], an open source tool which automatically annotates the mentions of DBpedia resources in a text and hence can be used to link a text to Wikipedia, to annotate our data sources, namely the Patentsview data set for USPTO patents and the arXiv data set for publications. To reduce the noise, we drop annotations which occurred less than 5 times. After that, we filter them by only keeping the above mentioned technologies. These annotations are then used to create time series, by summing up the number of times a technology ($t \in T$) has been mentioned in each year for each data source ($d \in D$). Because the number of patents and publications increase over time, we compute relative counts with respect to the total number of all technologies counts per year.

Furthermore, we obtain monthly Wikipedia page views of all the technologies to compute the impact score as described in section 4.2. The list of categories from Wikipedia was also extracted to compute the coherence score (see section 4.2). Finally, we aggregate and normalize these four scores to generate a final score for each technology.

People tend to create many articles in the Wikipedia (*Machine learning*, *Deep learning*, *Artificial Neural Network*) that are very close to each other. We use the Wikidata properties *subclass of*, *part of*, *instance of* or *said to be same as* to establish connection between such technologies. This allows us to join various technologies, thus creating the what we named a Technology class scoring.

4.1 Technology classification

The output of annotating abstracts from patents and publications is quite noisy. Each annotation refers to an article (not necessarily technology) in Wikipedia. Therefore, we developed a methodology to select technology articles from Wikipedia in order to keep only the relevant annotations for our use case. This methodology is a two-step process, where we first classify Wikipedia categories as technology/non-technology, followed by the association of articles to categories.

Each Wikipedia article is linked to categories, and these categories are linked to each other through a parent child relationship, forming a complex graph. In addition, Wikipedia also provides Main Topic Classifications (MTC), which include a list of 28 categories like Technology, Business, Arts, Health, etc. Even though Wikipedia contains a main topic named Technology, the edges between categories are very loosely defined (”is related to”). Therefore, we cannot rely on that concept to extract all technology articles.

To create a training data set for the classifier, we start cleaning up the directed categories graph by removing hidden categories, admin and user pages. This is followed by a set of regular expression filters, which remove categories referring to companies, people names, brands, currencies, countries, etc. Afterwards, we calculate the shortest path for each category in the filtered graph to the 28 MTC, keeping the ones with smallest distance to Technology, Science, or Engineering concepts. At each step of this process, the largest (weakly) connected component of the graph is retained. This results in a list of 7,876 categories which we manually labeled as *technology* or *non-technology.* Finally we end up with a list of 1,356 technology categories, while all other categories are considered non-technologies.

We use a combination of TF-IDF weighted bag of words and a vector representing the distance of the category to each MTC as the the input features. The abstracts of all pages directly linked to a category are concatenated and stemmed, followed by TF-IDF based weighting to generate a weighted bag of words for each category. This is followed by feature reduction to generate usable feature vectors. We observed best results using the mutual information based feature reduction with a target vector length of 1000. The distances to each topic in the MTC are appended to this vector to generate the final feature vectors.

Wikipedia contains about 1.4 million categories, while our training set has only 1,356 positive samples. Given such a big imbalance in class distribution, scaling and oversampling techniques are used while training the classifier. We experimented with different known algorithms and observed best results using Borderline-SMOTE[13] technique for oversampling followed by a SVM based classifier. This classifier is used to obtain a sub-graph of the Wikipedia categories graph, consisting of only technology categories.

Wikipedia article is considered a technology if it is directly connected to any category identified as a technology in the previous step. As a post-processing step, these articles are run through the regex based filters used in the first step. The final list is used to filter the annotations from patents and publications.

4.2 Scoring

Novelty score Scholars have defined radical novelty or less prior development as one of the key characteristics of an emerging technology [28, 10]. With regard to our data set, a radically new technology should have more mentions in recent years. If for a particular technology a large fraction of references occur in the last few years, it should get a high novelty score. To achieve this, we took the time-span of the last 8 years (in our case 2010-2017) and compute the percentage of annotations for each year. Linearly decreasing weights w_y from 8 to 1 were assigned to the years 2017 till 2010 respectively, thereby giving a higher weight to more recent years. Technologies for which most of the annotations occurred more than 8 years ago do not satisfy the novelty criterion and hence are discarded.

We start by defining yearly time series $X_{t,d}$

$$X_{t,d} = \{X_{t,d,y} : y \in Y\} \tag{1}$$

where

$X_{t,d,y}$ = number of times a technology occurred
$y \in Y$ = year in range (in our case 2010-2017)

Thus, the total number of occurrences of all technologies $t \in T$ in data set $d \in D$ over a given year y can be written as

$$Total(t, d) = \sum_{y \in Y} X_{t,d,y} \tag{2}$$

The novelty score $Novelty(t)$ of a technology $t \in T$ is then expressed as

$$Novelty(t) = \sum_{d \in D} \sum_{y \in Y} \Big(\frac{X_{t,d,y}}{Total(t, d)} \times 100 \times w_y \Big) \tag{3}$$

where w_y is the weight for each year and is given by

$$w_y = (y + 1 - \min_{\forall y' \in Y} y') \tag{4}$$

Growth score We deployed a two-step approach to compute the growth score of a technology. How fast a technology grows can be measured by the growth curves in patents and publications [28, 10, 1, 23].

In step 1, we utilize regression techniques to fit number of yearly technology mentions to 4 different curve models: linear, quadratic, Gaussian and exponential. We used the Apache Commons SimpleRegression and OLSMultipleLinearRegression for the linear and quadratic models. The same regression tools were used with the log of the data points to derive the exponential and Gaussian models, respectively.

Table 1. Curve fitting classes and growth scores

Class	Model_score
exponent increase/decreases	+/- 1.00
quadratic increase/decreases	+/- 0.75
Gaussian trend, increase/decreases	+/- 0.50
linear increase/decreases	+/- 0.25
Nothing fits (if $R^2 < 0.50$)	0.00

We choose the model with the highest R^2 statistical measure and compute the slope of the curve based on the regression coefficients. With the positive or negative sign of the slope we define the trend to be increasing or decreasing. Hence, on the basis of the best fitting model and the slope, we assign the technology to one of the classes defined in table 1 to compute the Model score.

In step 2, we compute the slope of the technology growth curve by taking the difference between the absolute counts of the last and the first year divided by the total number of years and normalize all values to the range [0.0; 1.0].

$$Slope(t, d) = \frac{Count(t, d, Y_{\text{final}}) - Count(t, d, Y_{\text{begin}})}{Y_{\text{final}} - Y_{\text{begin}}} \tag{5}$$

$$Norm\_slope(t,d) = \frac{Slope(t,d) - min(Slope(T,d))}{max(Slope(T,d)) - min(Slope(T,d))} \tag{6}$$

where:

$Norm\_slope(t,d)$ = normalized slope score of technology t in data set d
$Y_{\text{begin}}, Y_{\text{final}}$ = starting year, final year

The technology's final growth score is then computed from both the model and the slope score.

$$Growth(t) = \sum_{d \in D} \Big(Model\_score(t,d) + Norm\_slope(t,d) \Big) \tag{7}$$

Impact score We use the Wikipedia page view (w) statistics to compute the impact score of a technology. The Wikipedia API provides page view counts from July 2016 until present. Since the time span of less than 3 years is too small for yearly time series, we use a granularity of one monthly to get more data points.

After extracting the monthly views, we smooth the time series using a 3 months moving average filter and apply again the two-step approach as used for the growth score (section 4.2, with w representing the data set d). We classify the trends into the same 9 classes (Table 1), compute their slopes, and add the two scores to obtain a final score for the high impact criteria.

$$Impact(t) = Model\_score(t,w) + Norm\_Slope(t,w) \tag{8}$$

Coherence score We can label a technology as coherent if there's also a category with the same name in Wikipedia. The reasoning behind this is that the creation of a category manifests the branching out of a technology from its parent[27].

We use all unique categories of Wikipedia (Category_set), map the plural names to singular, and match them to articles with the same name. The coherence score is then computed with the following formula:

$$Coherence(t) = \begin{cases} 0.5, & \text{if t} \in \text{Category\_set} \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

Final score The final score is calculated by adding the normalized scores for novelty, growth, impact, and coherence and once again normalizing the sum to the range $[0.0; 1.0]$.

$$\begin{aligned} Final(t) = Norm[&n * Novelty(t) + g * Growth(t) \\ &+ i * Impact(t) + c * Coherence(t)] \end{aligned} \tag{10}$$

The values of parameters n, g, i and c are determined empirically with a goal of achieving the highest precision.

Technology class score We define as a technology class TC as a set of, by means of afford mentioned Wikidata properties, related technologies. The technology class is named as the top level technology within the set. The technology class score (TCs) is defined as follows:

$$TCs = \max_{t \in TC} Final(t) \tag{11}$$

5 Machine Learning

Machine Learning algorithms can infer mathematical relationships from simple data points, and this makes it a viable alternative to the first application method. Using the US Patent Office database, which contains standardised metadata on over 7,000,000 patents, we have a solid and extensive data source from which we can train several supervised machine learning classifiers. Supervised algorithms train on a subset of the original data set, and following this their predictive ability is tested on the remaining subset. Classifiers don't map their results onto a continuous output, such as in Equation 1, but rather classify each data point into one of N categories.

6 Observations

6.1 Results from scoring algorithms

In this section, we present some observations from various intermediary steps for scoring emerging technologies followed by observations from the final scoring of technologies.

Column Technology in the Table 2 shows the top 20 technologies according to the final score. For the results presented in this table the parameters n, g, i and c within the equation 10 were set to 0.6, 0.7, 0.1, 0.2 respectively (we denote this parameter set as max_prec). *Deep learning* is the top trending technology by our measures. We find *Convolutional neural network* a sub-categories of *Deep learning* also in the list. Figure **??** shows relative trends for the top technology *Deep learning* within the Arxiv and Patentsview deatsets. We can clearly see strongly increasing trend in both data sources. As expected, *Machine learning* appears in the list, as does the *Internet of things* both being coherent and also showing up in top 20 technologies in terms of impact and novelty, respectively. The *Cyberattack*, appearing high in the list, as well as various other technologies related to *Computer security*, forming the second group of technologies in this result list. *Key-value database* – the simplest form of NoSQL databases – appears seventh in the top 20 emerging technologies.

Our algorithm suggests the emergence of other technologies like *Cloud gaming*, *Communication* and *Smartphone*, which have received attention since years. We also notice currently sought-after technologies like *Autonomous car*, *Knowledge graph* and *5G* in the top 20 scored technologies. Taking the approach presented in [32] we can argue that our algorithm returns 4 convergence emerging

Table 2. Top 20 technologies and Technology classes

Technology	**Technology class**
Deep learning†	Artificial intelligence
Autonomous car	Autonomous driving
Internet of things	Internet of thing
CNN†	Computer security
Machine learning†	Database
Ransomware∗	Knowledge Graph
Key-value database	Augmented, virtual, mixed reality
Shard (database architecture)	Connectivity
Cyberattack∗	Telecommunication
Knowledge graph	Cloud and virtualization
Augmented reality	Data Science
Smartphone	Optical instrument
Communication	Virtual assistant
Side-channel attack∗	Exoskeleton
Cloud gaming	Computer vision
5G	Satellite imagery
Data science	Heterogeneous computing
Return oriented programming	Distributed computing
Lidar	Medical device
Push technology	3D printing

technologies (CET) in top 5 results, with the fifth (CNN) being a sub-class of Deep Learning.

Technology classification The technology classification algorithm described in section 4.1 recognized 50,954 technologies from 4,996,310 articles in total and 2,996 technology categories from a total of 1,481,291 categories from the English Wikipedia. For the evaluation of our methodology, we held out 10% randomly selected samples from the complete data. The holdout test set was selected before oversampling with the same class ratios as the complete data set to effectively simulate a real world use case. The training was performed using cross validation on the remaining data. The model achieved a True Positive Rate of 89%, a False Positive Rate of 0.2% and a precision of 42%. The low precision is due to the highly imbalanced nature of the dataset, with very few positive examples of only 1%. This model was used to generate the list of technologies, which was used to filter the annotations from arXiv and patents. 15,259 unique technologies from this list were finally mentioned in the two sources, with 1,916 of them having the corresponding category, thus considered to be coherent. We evaluated manually 100 top scoring "technologies" in order to estimate the precision of this binary classifier. Among those 100 results we have found 15 Wikipedia articles incorrectly classified as technology without being one.

Emergence scoring We evaluated these results for average precision and recall, using the evaluation set presented in the following section as the ground truth.

Table 3. Evaluation set, Technology classes based on emerging technologies proposed by analyst (Gartner, Forrester, IHS Markit and WEF)

Technology class	
Tissue engineering	Unmanned aerial vehicle
Smartdust	Artificial Intelligence
4D printing∗	Ontology(information science)
Neuromorphic engineering	Exoskeleton
Edge computing	Autonomous driving
Self healing system technology†	Volumetric display
5G	Quantum computing
Platform as a service∗	Application specific Integrated Circuits
Autonomous Robot	Mobile Robot
Brain Computer Interface	Internet of Things
Biochip	Digital twin
Nanotechnology	Virtual assistant
Lithium-silicon battery	Blockchain
Augmented, Virtual, Mixed reality	E-textiles
Cloud computing	Computer vision
Ubiquitous Video†	Natural Language Generation∗
Switched fabric	Personalized medicine
Cell encapsulation	Gene drive

For evaluating the correctness of our algorithm, we took the union set of emerging technologies published by the leading technology analysts Gartner, Forrester, IHS Markit, and the World Economic Forum (WEF) from 2018. For this year, Gartner predicted 35 technologies in its technology hype cycle. Forrester predicted 12, IHS Markit 8, and WEF 10 emerging technologies.

Merging the overlapping technologies from these four list resulted in a list of 36 unique technology classes. In this case the same approach was used as for the technology classes within the result set. Table 3 gives an overview of these classes. A strong bias toward Computer science can be noticed in the data presented in these tables, with 72% of technologies being linked to it. In this table we have marked with ”†” technologies we were unable to map directly to a Wikipedia article/category. Finally, the articles judged to be non-technologies by the classifier are marked in the tables by ”∗”. The Augmented, Mixed and Virtual reality Wikipedia articles are presented together while being proposed as a single technology by Forrester.

We have chosen to evaluate the performance of our system on the first list of top 20 returned results both in terms of technology and technology class using 3 distinct set of parameters in the equation 10. We have chosen this cutoff for our results, taking into account that search engine users rarely go beyond second page of results[7]. We consider as relevant any technology/class in our result lists linked to a technology illustrated in the Table 3 by means of any of previously mentioned Wikidata properties. In the "base" run, all parameters in the equation are set to 1. In addition to the "max_prec" parameter set we also present the average precision and recall of the computer science technology class (max_prec_cs) with parameters being chosen to facilitate the max precision in this particular setting (e.g g, n, i and c set to 1, 0.3, 0.1 and 0.3 respectively). Within the 20 technologies with the highest final score, we can observe only one non-technology result. The average precision (AP) calculated for these results (Table 4), based on the previously described relevance judgment, elevates to 0.72 for the base run. However, all the relevant concepts from this sub-set relate to only 6 out of 36 technologies mentioned before, recall (R) of 0.16. Using the max_prec parameter set we were able to augment both AP(0.81) and R(0.19). For the same set of parameters, row TC gives the average precision and recall for technology classes, where we can notice an important increase in the recall.

Table 4. Average Precision (AP) and Recall (R) of the technologies (Tech) and technology classes (TC)

Parameters	Classes	AP	R
base	Tech	0.72	0.16
max_prec	Tech	0.81	0.19
	TC	0.72	0.28
	CS TC	0.79	0.36
max_prec_cs	CS TC	0.90	0.36

In overall results returned by the system, we can observe the same bias concerning computer science we have noticed within our evaluation set, with 70% of technologies within the top 100 results belonging to this domain. The same bias is observable in data sources both those used in this work as well as in other sources (ex. Microsoft Academic Graph[8]). This bias makes it difficult to explore trends in other domains. If we take an example in chemistry, the International Union of Pure and Applied Chemistry (IUPAC) has issued the list of emerging technologies[9] for this domain containing among others *3D-bioprinting* or *Flow chemistry*, none of which figure in our evaluation set, while being present in our technology result set, ranked 4897 and 12421 respectively.

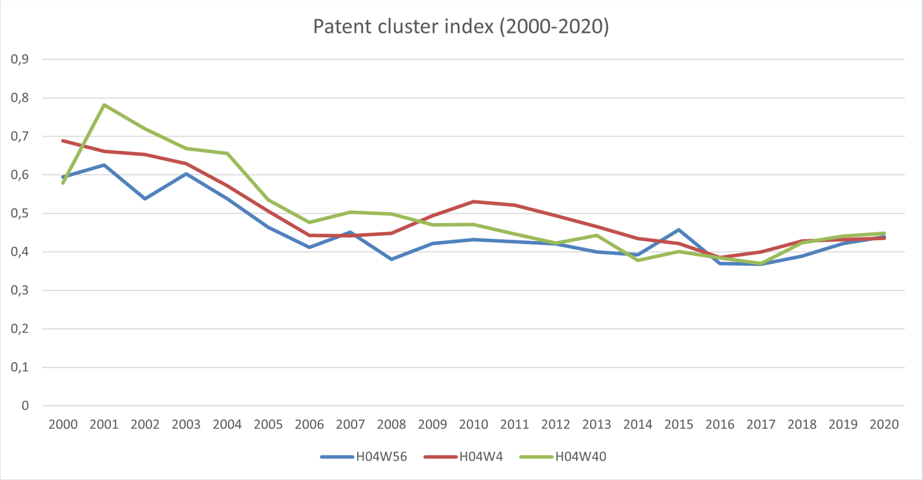
Splitting, in addition to the technology classes result set creation, the result set as well as the evaluation set into distinct domains (CS, Nanotechnology, Medicine etc.) allows us to work around above-mentioned bias. The third row (CS TC) of the Table 4 gives the average precision and recall when only results related this field are being taken into account, this class being predominant in our result/evaluation sets. Even though this approach results in only 10% increase in the average precision, the increase in recall elevates to 30%. Exploring this bias further we have been able to determine the set of parameters (max_prec_cs) capable in achieving the precision of 0.9 in computer science related class setting.

[7] https://www.forbes.com/sites/forbesagencycouncil/2017/10/30/the-value-of-search-results-rankings/#60410ae544d3

[8] https://academic.microsoft.com/home

[9] https://iupac.org/what-we-do/top-ten/

6.2 Results from classifier



7 Conclusion, Limitations, Future scope

This paper describes an automated method for identifying emerging technologies based on publicly available data. Our method is neither limited to any particular technology sector, nor does it involve human subject matter experts. All results presented in this paper have a clear mathematical foundation and can be reproduced by any interested party.

We crated an individual emergence scoring based on novelty, growth, impact, and coherence scores. Scores for novelty and growth were computed from DBpedia annotated USPO patents and arxiv publications time series. Impact score was calculated Wikipedia page views time series, while the coherence score was derived from Wikipedia categories. Using an aggregated final score we produced a ranked emerging technologies list. Using the Wikidata properties, we elaborated an additional technology classes ranked list. A third results set was elaborated by eliminating all non Computer science technology classes from the technology classes list.

To evaluate these results we have elaborated an evaluation set of 36 emerging technologies by merging the lists identified by the market analysts Gartner, Forrester, IHS Markit, and WEF. Evaluation of 20 top scoring technologies revealed a very low recall (0.16). By adjusting the parameters in the final score calculation, we were able obtained an increase in both precision and recall. However using the technology classes result set we obtain an important increase in recall (0.36), without significantly hurting the average precision (−11%).

In result sets, evaluation set and data sources used in this study the CS bias is very important. By domain wise splitting of results and evaluation sets and focusing on the CS related technologies, we have obtained an important improvement in both average precision and recall (10% and 30% respectively).

Finding the first Nanotechnology (only Chemistry related class in evaluation set) related technology on 51 place in our result list, confirms that this is the adequate manner for non CS technologies to become visible. Including additional domain specific sources might also help in overcoming this problem. This would also imply using additional domain specific evaluation sets, as well domain specific parameter setting accordingly.

Our algorithm detects certain technologies receiving an important attention lately such as *Heterogeneous computing* not being present in the evaluation set. In addition, *Computer security* related technologies take an important place in our results. Even thought these technologies might not be emergent, they remain highly important.

Using this research as a starting point we would like, in the future, to elaborate a methodology to determine in what stage of emergence (pre-emergence/ emergence/ post-emergence), if any, is a particular technology in the given time frame.

References

1. B. Andersen. The hunt for s-shaped growth paths in technological innovation: a patent study. *Journal of evolutionary economics*, 9(4):487–526, 1999.
2. M. Bengisu and R. Nekhili. Forecasting emerging technologies with the aid of science and technology databases. *Technological Forecasting and Social Change*, 73(7):835 – 844, 2006.
3. W. Boon and E. Moors. Exploring emerging technologies using metaphors: A study of orphan drugs and pharmacogenomics. *Social Science & Medicine*, 66(9):1915 – 1927, 2008.
4. M. Borup, N. Brown, K. Konrad, and H. Van Lente. The sociology of expectations in science and technology. *Technology analysis & strategic management*, 18(3-4):285–298, 2006.
5. Y. Cha and C. A. Stow. Mining web-based data to assess public response to environmental events. *Environmental pollution*, 198:97–99, 2015.
6. H. Choi and H. Varian. Predicting the present with google trends. *Economic Record*, 88:2–9, 2012.
7. N. Corrocher, F. Malerba, and F. Montobbio. The emergence of new technologies in the ict field: main actors, geographical distribution and knowledge sources. Economics and quantitative methods, Department of Economics, University of Insubria, 2003.
8. S. Cozzens, S. Gatchair, J. Kang, K.-S. Kim, H. J. Lee, G. Ordóñez, and A. Porter. Emerging technologies: quantitative identification and measurement. *Technology Analysis & Strategic Management*, 22(3):361–376, 2010.
9. T. U. Daim, G. Rueda, H. Martin, and P. Gerdsri. Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change*, 73(8):981 – 1012, 2006. Tech Mining: Exploiting Science and Technology Information Resources.
10. G. S. Day and P. J. H. Schoemaker. Avoiding the pitfalls of emerging technologies. *California Management Review*, 42(2):8–33, 2000.
11. W. Glänzel and B. Thijs. Using 'core documents' for detecting and labelling new emerging topics. *Scientometrics*, 91(2):399–416, May 2012.

12. M. Halaweh. Emerging technology: What is it? *Journal of Technology Management & Innovation*, 8(3):108–115, 2013.
13. H. Han, W.-Y. Wang, and B.-H. Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International Conference on Intelligent Computing*, pages 878–887. Springer, 2005.
14. R. Haupt, M. Kloyer, and M. Lange. Patent indicators for the technology life cycle development. *Research Policy*, 36(3):387–398, 2007.
15. S.-C. Hung and Y.-Y. Chu. Stimulating new industries from emerging technologies: challenges for the public sector. *Technovation*, 26(1):104 – 110, 2006.
16. G. Intepe and T. Koc. The use of s curves in technology forecasting and its application on 3d tv technology. *International Journal of Industrial and Manufacturing Engineering*, 6(11), 2012.
17. M. Kämpf, E. Tessenow, D. Y. Kenett, and J. W. Kantelhardt. The detection of emerging trends using wikipedia traffic data and context networks. *PloS one*, 10(12):e0141892, 2015.
18. D. Kucharavy, E. Schenk, and R. D. Guio. Long-run forecasting of emerging technologies with logistic models and growth of knowledge. In *Proceedings of the 19th CIRP Design Conference - Competitive Design, Cranfield University, 30-31 March, 2009*, page 277, 2009.
19. M. N. Kyebambe, G. Cheng, Y. Huang, C. He, and Z. Zhang. Forecasting emerging technologies: A supervised learning approach through patent analysis. *Technological Forecasting and Social Change*, 125:236 – 244, 2017.
20. C. Lee, O. Kwon, and D. Kim, Myeongjung an Kwon. Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technological Forecasting and Social Change*, 127(3):291–303, 2018.
21. B. R. Martin. Foresight in science and technology. *Technology Analysis & Strategic Management*, 7(2):139–168, 1995.
22. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8. ACM, 2011.
23. M. Meyer. Patent citation analysis in a novel field of technology: An exploration of nano-science and nano-technology. *Scientometrics*, 51(1):163–183, 2001.
24. M. Nieto, F. Lopéz, and F. Cruz. Performance analysis of technology using the s curve model: the case of digital signal processing (dsp) technologies. *Technovation*, 18(6):439 – 457, 1998.
25. A. L. Porter, J. D. Roessner, X.-Y. Jin, and N. C. Newman. Measuring national 'emerging technology'capabilities. *Science and Public Policy*, 29(3):189–200, 2002.
26. S. Ranaei, M. Karvonen, A. Suominen, and T. Kässi. Forecasting emerging technologies of low emission vehicle. In *Proceedings of PICMET '14 Conference: Portland International Center for Management of Engineering and Technology; Infrastructure and Service Integration*, pages 2924–2937, 2014.
27. D. Rotolo, D. Hicks, and B. Martin. What is an emerging technology? *Research Policy*, 44(10):1827–1843, 2015.
28. H. Small, K. W. Boyack, and R. Klavans. Identifying emerging topics in science and technology. *Research Policy*, 43(8):1450 – 1467, 2014.
29. B. C. Stahl. What does the future hold? a critical view of emerging information and communication technologies and their social consequences. In M. Chiasson, O. Henfridsson, H. Karsten, and J. I. DeGross, editors, *Researching the Future in Information Systems*, pages 59–76, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

30. M. Steinert and L. Leifer. Scrutinizing gartner's hype cycle approach. In *Technology Management for Global Economic Growth (PICMET), 2010 Proceedings of PICMET'10:*, pages 1–13. IEEE, 2010.
31. S. Telfer and J. Woodburn. Let me google that for you: a time series analysis of seasonality in internet search trends for terms related to foot and ankle pain. *Journal of foot and ankle research*, 8(1):27, 2015.
32. Y. Zhou, F. Dong, Z. Li, J. Du, Y. Liu, and L. Zhang. Forecasting emerging technologies with deep learning and data augmentation: convergence emerging technologies vs non-convergence emerging technologies. In *Proceedings of 6th International Conference on Future-Oriented Technology Analysis (FTA)*, 2018.