Scientometrics (2017) 111:743–772
DOI 10.1007/s11192-017-2268-3

A patent search strategy based on machine learning for the emerging field of service robotics

Florian Kreuchauff[2] · **Vladimir Korzinov**[1]

Received: 6 March 2016 / Published online: 10 February 2017

Abstract Emerging technologies are often not part of any official industry, patent or trademark classification systems. Thus, delineating boundaries to measure their early development stage is a nontrivial task. This paper is aimed to present a methodology to automatically classify patents concerning service robots. We introduce a synergy of a traditional technology identification process, namely keyword extraction and verification by an expert community, with a machine learning algorithm. The result is a novel possibility to allocate patents which (1) reduces expert bias regarding vested interests on lexical query methods, (2) avoids problems with citation approaches, and (3) facilitates evolutionary changes. Based upon a small core set of worldwide service robotics patent applications, we derive apt n-gram frequency vectors and train a support vector machine, relying only on titles, abstracts, and IPC categorization of each document. Altering the utilized Kernel functions and respective parameters, we reach a recall level of 83% and precision level of 85%.

Keywords Service robotics · Search strategy · Patent query · Data mining · Machine learning · Support vector machine

JEL Classification C02 · C18 · C45

✉ Vladimir Korzinov
vladimir.korzinov@kit.edu

Florian Kreuchauff
florian.kreuchauff@kit.edu

[1] Karlsruhe Institute of Technology, Rüppurrer Str. 1a, Haus B, 76137 Karlsruhe, Germany

[2] Geschäftsstelle Expertenkommission Forschung und Innovation (EFI) c/o SV Gemeinnützige Gesellschaft für Wissenschaftsstatistik mbH, Pariser Platz 6, D-10117 Berlin, Germany

Introduction

Innovation policies that address promising emerging technologies serve to reach macroeconomic objectives such as promoting sustainable growth and prosperity. They are legitimate due to the various uncertainties associated with new technological fields that result from coordination problems in complex innovation chains with scale economies, multilateral dependencies, and externalities. In order to develop effective policy measures, one has to carefully recognize emergence patterns and assess possible downstream effects. This is a demanding task since these patterns vary across technologies, time, scale, and regional and institutional environments. It is important that policy advices rely on credible data sources that aptly display early research and innovation results at the very beginning of value creation. However, as long as a new technology has not yet been specified within official statistical schemes, the identification of delineating boundaries in respective data bases is a nontrivial problem.

Service robotics (hereafter referred to as SR) is a current example of an emerging technology. Although robots have been in use in industrial production for several decades, services outside of manufacturing proved to be too complex in the past to be provided by a robotic machine. Instead, robots were confined to repetitive and monotonous tasks within production processes that demanded for high degrees of precision, strength, and endurance. In the last years, however, the range of possible applications has started to expand. Decreased computing times and increased data storage capacities have laid the foundations for multidimensional sensory perception. Moreover, new machine learning tools have facilitated robots in becoming more adaptive, enabling them to cope with unstructured environments, human interaction, and instantly changing requirements. In industrial robotics (hereafter referred to as IR), we already see some prototypes of collaborative lightweight robots. These are much cheaper than their old-fashioned hulking predecessors, although far easier to operate and reconfigure, opening new areas of application where manual skills have been labeled 'irreplaceable'—for instance regarding the assembly of small components with low batch sizes in microelectronics. Human operators direct these robots using voice and gesture control.

More generally speaking, latest developments in man–machine interaction (MMI) make it possible to complement human labor with an increasingly efficient, yet easily controllable machine environment. Robots are leaving the strictly fenced security zones that are common to protect the human workforce in IR. These more service-oriented robots clean buildings, provide support during surgical procedures or assist in dangerous maintenance and inspection work.

The International Federation of Robotics (IFR) has been working on a service robot definition and classification scheme since 1995. A preliminary definition states that a service robot is a robot that performs useful tasks for humans or equipment excluding industrial automation applications. Industrial automation applications include, but are not limited to, manufacturing, inspection, packaging, and assembly (compare www.ifr.org and ISO 8373:2012). Service robots can be further subdivided into those for noncommercial personal use like domestic servant robots or automated wheelchairs, and those for professional commercial services, for which they are usually run by trained operators like firefighting or surgery systems in hospitals. Hence, SR contribute to both traditional and a variety of new types of services.[1]

[1] Beyond its potential productivity effects, SR is believed to induce visible changes in employment structures (Autor et al. 2003; Frey and Osborne 2017; Graetz and Michaels 2015). Its potential to change organization processes in firms as well as everyday life of people is already visible in the diffusion of semi-autonomous physical systems out of industrial fabrication and into service economies.

Apart from its multiple applications, SR itself consists of various technologies. Like many new emerging fields, it lies on the crossroads of many disciplines such as mechanics, electronics, control systems, informatics, and others. Economics already has made theoretical attempts to model this combinatorial nature of new technologies (Arthur and Polak 2006) which essentially embody economic arrangements within a complex ever changing system (Arthur 1999, 2009). Due to this complexity, it is hard to disentangle new technologies from their parts at early stages of development. As a result of the arising multiplicity, SR is so far not clearly confined in databases and thus neither part of any existing official industry, patent or trademark classification system nor of any concordances, not to mention national account systems. Therefore, distinguishing SR from IR within such data bases is hardly possible. This so far has impeded a comprehensive assessment of the economic impacts of SR diffusion, especially with respect to the magnitude, timing, and geographical localization.

With our work, we make SR tractable by developing a search strategy to identify it within patent databases. Moreover, we model the approach not to be limited to patents but to be applicable to scientific publications as well. In addition, the general methodology is not even confined to the field of robotics, but could be applied to any similar identification problem. Differentiating from classical lexical and citation approaches used by other scholars, our approach introduces a machine learning algorithm that is utilized as a classifier. Being trained on some sample data, this classifier acts as an 'expert'. With a certain degree of precision, the machine is able to decide whether a patent belongs to the category of service robotics or not. Since there are several approaches in the scientific literature which deal with analogous problems of technology detection and classification, we hereby set out to (1) limit expert bias regarding vested interests on lexical query methods (with respect to term inclusion and exclusion), (2) avoid problems with citation approaches such as the lack of portability, and (3) facilitate evolutionary changes.

The following sections are organized as follows: first, we give an overview of previous technology identification approaches referring to examples of similar emerging fields that lacked classification schemes in their infant phase. Second, we present our step-by-step methodology for identifying developments in an emerging field characterized only by its early applications. It successively describes the use of patents as apt data sources, the retrieval of a structured core dataset, and the use of an automated machine learning algorithm, namely a support vector machine (hereafter referred to as SVM). Finally, we present results of our pioneering approach and conclude with the future scope for improvement.

Literature review

There is no widely-agreed upon definition of emerging technologies (Halaweh 2013). The initial lack of common knowledge, standards, and specifications entails uncertainties along various dimensions (Stahl 2011). Future costs and benefits, relevant actors, adoption behavior, and potential socio-economic implications such as creative destruction are highly unclear (Srinivasan 2008). Therefore, scientific studies have been using bibliometrics and patent analysis to monitor trends for a variety of domains and assess the nature of emerging technologies already within scientific research and early development.[2] We will argue later on, why patents are an apt choice for our purpose.

[2] Cozzens et al. (2010) argue that bibliometric data, in particular proposals and publications, seem to be most useful for monitoring the technological horizon. Patent analysis on the other hand, besides being long known to be valuable for competitive and trend analysis (Abraham and Morita 2001; Liu and Shyu 1997), has become sophisticated to even predict emerging fields (Erdi et al. 2013).

No matter what the paramount aim, all analyses greatly rely on well-founded data acquisition, which first and foremost identifies the technology under consideration. The first-best approach in this regard relies on existing classifications. For instance, a recent study on behalf of the European Commission (cf. Frietsch 2015) makes use of the so-called WIPO patent classification following Schmoch (2008) regarding technology concordances, and of another existing classification provided by Van de Velde et al. (2013) for so-called Key Enabling Technologies (KETs). Moreover, the study addresses Societal Grand Challenges identified as priority areas in Horizon 2020 and Europe 2020. For some of the technologies within these areas, for instance biotechnology, ICT, or environmental technologies, the OECD provides definitions.[3] Moreover, some patent search strategies and technology definitions are provided in annual reports of various patent offices, for example on electric and hybrid vehicle technologies, renewable energy technologies, or biotechnology.[4] For all other (emerging) technology (sub-)fields that are considered relevant, the study develops patent classifications itself—with considerable effort.

The most striking example of the past concerns nanotechnology. Early conceptions of apt queries for nanotechnology proved to be difficult, as the first specific patent class within the International Patent Classification (IPC), subclass B82B,[5] which basically refers to nanostructures and their fabrication, was not introduced before the year 2000 and did not incorporate applications from former years (Noyons et al. 2003). In its infancy, it contained only an estimated 10% of all relevant documents. Hence, the first scientific identification approach for nanoscience and technology relied instead on a keyword-based (or simply *lexical*) query developed in 2001 by Fraunhofer Institute for Systems and Innovation Research (ISI) in Germany and the Centre for Science and Technology Studies (CWTS) at Leiden University in the Netherlands—again, with considerable effort.

Whenever classification schemes are missing or impossible to develop, technology identification within patent or publication databases has to be predicated on either (1) lexical, (2) citationist, or mixed search strategies.[6] A lexical query (1) is a search for specified terms, which in the most simple case might consist of only one word (like 'nano*' for nanotechnologies) or a basic combination (like, in our case, 'service robot*'). This primal string is applied to titles, abstracts, keywords or even the whole text body of examined documents. Some of these documents might prove to be relevant in the eyes of experts and thus offer additional terms starting an iterative process.[7] Considering emerging fields, the number of terms within a search string that is developed in such a lexical manner naturally grows rapidly. More and more scholars and practitioners become attracted by the

[3] Cf. http://www.oecd.org/sti/intellectual-property-statistics-and-analysis.htm#method.

[4] See for example the annual reports by the German Patent and Trademark Office at http://www.dpma.de/english/service/publications/annualreports/index.html.

[5] Only in 2011, a second subclass, B82Y, focusing on specific uses or applications of nanostructures was introduced for IPC and the Cooperative Patent Classification (CPC). Previously, related nano patent documents could only be identified if they were classified via the European Classification System (ECLA) with the subclass Y01. ECLA *Y*-codes have been created as an extension of the original classification system, to extend classification capabilities to new (emerging) technology areas of special interest.

[6] With respect to scientific publications, another common strategy is to identify core journals. All articles within those journals are then considered relevant. For patents though, this search strategy is obviously not feasible, which is why we do not deepen it any further.

[7] Such a search strategy is called evolutionary, if subsequent researchers may build upon existing query structures by progressively incorporating terms that better specify the technology and widen its scope (Mogoutov and Kahane 2007).

field,[8] adding alternatives and broadening interpretations in the course of time. Referring to nanotechnology as a striking example again, in order to keep track of the dynamically spreading nanofields, Porter et al. (2008) discussed a modular Boolean keyword search strategy with multiple-step inclusion and exclusion processes, which had to be subsequently enhanced and evolutionarily revised (Arora et al. 2013). Identification problems are heightened by the fact that both authors of scientific publications as well as applicants of patents are interested in some rephrasing: The former, because they might benefit from a serendipity effect if their label establishes itself in the scientific community. And the latter because of encryption and legalese issues: Applicants may want to relabel critical terms, both to hide relevant documents and technical information from actual rivals and to build patent thickets of overlapping IPR which precludes potential competitors from commercializing new technology altogether.

A lexical query can be enriched (or fully substituted, if a core of documents is already verified) by adding documents and inherent terms identified via citation approaches (2), for instance by including new publications that are cited by at least two authors belonging to an initial database (Garfield 1967; Bassecoulard et al. 2007)[9] or, regarding patents, by including applications that refer to prior art that has been part of the previously established core. In our example, Mogoutov and Kahane (2007) enriched an initial nanostring by a number of subfields, automatically identified and defined through the journal intercitation network density displayed in the initial core dataset of nanodocuments. Relevant keywords linked to each subfield were then tested for their specificity and relevance before being sequentially incorporated to build a final query.

The instance of nanotechnology illustrates well how much effort is involved in the development of an evolutionary query. Lately, private interests—rather than governmental or scientific research—have driven even more elaborate technology identification procedures, making use of ongoing technological advancements as well as computational power: Companies that seek to monitor competitors or investigate latest research trends have started to rely on more cost-efficient processes in order to lower resulting expenditures. As a side effect, some encompassing literature on specialized text mining techniques has emerged, which goes beyond lexical and citation-based procedures. To name just a few, Li et al. (2009) attempt to find significant rare keywords considering heterogeneous terms used by assignees, attorneys, and inventors. Yoon and Park (2004) argue that citation analysis has some crucial drawbacks and propose a network-based analysis as alternative method that groups patents according to their keyword distances. Lee (2008) uses co-word analyses regarding term association strength and provides indicators and visualization methods to measure latest research trends. Lee et al. (2009) transform patent documents into structured data to identify keyword vectors, which they boil down to principal components for a low-dimensional mapping. These facilitate the identification of areas with low patent density, which are interpreted as vacancies and thus chances for further technical exploitation. Erdi et al. (2013) use methods of citation and social network analysis, cluster generation, and trend analysis. Tseng et al. (2007) attempt to develop a holistic process for creating final patent maps for topic analyses and other tasks such as patent

[8] For the instance of nanotechnology, to which we refer throughout, Arora et al. (2014) measure the growth in nano-prefixed terms in scholarly publications and find that the percentage of articles using a nano-prefixed term has increased from less than 10% in the early 1990s to almost 80% by 2010.

[9] This approach naturally harbors the risk of including generic articles of any scientific field that somehow happen to be cited in a technologically unrelated context. Bassecoulard et al. (2007) therefore incorporate a statistical relevance limit relying on the specificity of citations.

classification, organization, knowledge sharing, and prior art searches. They describe a series of techniques including text segmentation, summary extraction, feature selection, term association, cluster generation, topic identification, and information mapping. Engineering research itself shares some interest in following latest developments as well. For the field of robotics, Ruffaldi et al. (2010) is a good instance: They visualize trends in the domains of rehabilitation and surgical robotics identified via text mining.

SVM and other methods of supervised learning were extensively used in social science research for a variety of purposes. McKeown et al. (2016) used it for prediction of promising scientific concepts in publications. Lee et al. (2010) looked at the probability of network tie formations. Another application was for prediction of citation counts made by Fu and Aliferis (2010). Supervised learning is widely used in various methods for topic modeling by Lee et al. (2010). Finally, it was used for classification for example of academic web pages by Kenekayoro et al. (2014). All these papers feature either different focuses or applications of machine learning methods. Our main contribution lies in the new field of study (SR) and application to patent data.

Methodology

Following Mogoutov and Kahane (2007), the relative performance of different identification approaches may be compared via the respective degree of intervention of experts, their portability, their transparency regarding core features and respective impacts on final results, their replicability, their adaptability, meaning their ability to produce valid results while the technology in question keeps evolving, their updating capacity, and the extent and relevance of the data obtained. Certainly, no single best approach exists, since any method has its advantages and drawbacks according to these criteria. We will conclude on the relative performance of our approach at the end of this paper.

In line with the current text mining literature, we propose a machine learning algorithm instead of a purely lexical, purely citationist or mixed query. Consequently, we first identify a small core patent dataset consisting of 228 patent applications and then let automated algorithms identify emerging technology boarders.

Patents as data source

As soon as a technology is sufficiently well specified, generically distinguishable, and ideally properly classified, there are various techniques to map ongoing advancements. However, if such a delineation is not yet established and no broadly accepted consensus has been reached so far, economists most often rely on lexical, citation-based, or mixed search strategies for prior identification purposes that help to trace related emerging fundamental and application knowledge in academic articles and patent documents.[10] As regards the technology under consideration, it is important to acknowledge that according to the IFR, the intended use, and as a consequence, the factual field of application determines the delimitation of SR from IR. Thus, patents are the data source of choice for an automated SR identification, since patentability requires an indication of the intended commercial implementation. Patents, despite all difficulties that arise in

[10] Consequently, the adequate data sources for this identification process are the same that comprise the targets of subsequent analyses which might give cause for some criticism.

their use and interpretation,[11] are widely accepted as indicators of innovative activity (Griliches 1990; Hall et al. 2005). Especially citation structures facilitate tracing knowledge flows (see, for instance, Jaffe et al. 1993; Thompson 2006; Fischer et al. 2009; Bresnahan 2010) and thus make technology development patterns visible. Hence, we started with a patent search strategy with a vision to extrapolate it to other lexical sources.

Building a structured core dataset that is suited for later application in machine learning requires the identification of a sufficiently large number of documents that are validated as part of the technology and capture most of its hitherto variety of developments. This validation is granted by independent technological experts, who can either provide those documents themselves or may be given a predefined assortment to adjudicate on. The latter decreases a potential expert bias with respect to multifaceted preferences but might give rise to a negative influence of the researcher himself, who has to develop a search method for this primal assortment. Facing this trade-off, we decided to provide experts with a predefined core dataset.

Retrieval of a core SR patent dataset

All unstructured patent text data as well as related document metadata were extracted from the 'EPO Worldwide Patent Statistical Database' (PATSTAT), version as of April 2013.[12] First, we extracted all patents that were either sorted in IPC class B25 J[13] or contained a substring like 'robot*' in their respective titles or abstracts.[14] Hence, we established a set of documents describing robotic devices. Second, in order to identify a subset of potential SR patent documents that comprise most of the hitherto existing developments, we created 11 subqueries based mainly upon IFR application fields for service robots. These queries consisted both of IPC subclasses (mostly at a four-digit level) and stemmed lexical terms, combined modularly in a Boolean structure.[15]

The second step provided us with 11 subsamples of potential SR patents. While other approaches regarding similar tasks of technology identification from there on further evaluate candidate terms by testing, assessing and adjusting terms and class codes to address weaknesses and follow emerging research trails manually (Porter et al. 2008), we did not alter the primal modular Boolean search. This allowed us to develop a dataset that would not compromise our analysis and would include all potential SR patents. As indicated above, we left the verification of the underlying categorization to technological experts. Two independent academic expert groups with 15 scientists, affiliated with the

[11] For the technology under consideration in this paper, it is important to note that SR patents are very much different from business process and service patents and there is little if no overlap. SR patents are much more 'technological' in the sense that they contain information about how a robot is constructed and for which environment its functionalities are intended. Their content is thus close to IR patents, making them hard to disentangle from each other. In contrast, business and service patents contain organizational innovations to a large extent.

[12] This database encompasses raw data from about 60 million patent applications and 30 million granted patents, utility models, PCT applications, etc. filed at more than 100 patent authorities worldwide.

[13] Manipulators; Chambers Provided With Manipulation Devices. See http://www.wipo.int/classifications/ipc/en/.

[14] According to the USPTO, most of the manipulators classified in B25 J are industrial robots. See http://www.uspto.gov/web/patents/classification/cpc/html/defB25J.html.

[15] We have included one example of such a subquery in the appendix. All other queries are available upon request.

- High Performance Humanoid Technologies (H2T) from the Institute for Anthropomatics and Robotics at KIT, Germany, and the
- Delft Center for Systems and Control/Robotics Institute at TU Delft, Netherlands

took on the task to decide which of the patents belonged to SR and which belonged complementarily to IR. The above experts were specialized in humanoid robotics, computer science, and mechanical engineering. Their experience in the field of robotics varied between 1 and 15 years.

We provided them with 228 full-body versions of potential SR patents from all over the world, which we had identified via our primal subsample queries within PATSTAT. These full-body patents had to be manually researched, since PATSTAT only offers English titles and abstracts for text mining purposes. Thus, the judging scientists could not only refer to these text parts but as well to all engineering drawings. In doing so, we intended to make the initial judgment as robust as possible. Furthermore, our experts could rely on the patent's claims in some cases—as long as these claims were available in English. However, this was not necessarily the case, since PATSTAT also lists patent applications from various international offices that do not translate claims.

For the application of automated machine learning approaches, we then transformed the unstructured patent document text available in PATSTAT into structured data. It is important to note again that for our machine learning purposes, we hereby could not rely on patent claims with regard to their absence in PATSTAT. One has to point out that adding claims to the analyzed text bodies, the resulting recall and precision of the procedure might be further improved, given a certain computing power to cope with the exponentially growing number of data to be processed. However, even without claims included we set out to demonstrate the validity of the identification method shown in the following sections.

The transformation into structured data included several steps, namely (1) combining titles and abstracts in one body and splitting the resulting strings into single terms in normal lower cases, (2) removing stop words, (3) stemming, i.e. reducing inflected words to their stem, (4) constructing n-grams of term combinations (up to 3 words in one), and (5) deriving normalized word and n-gram frequencies for each document.[16]

With these normalized frequencies, a matrix was constructed with columns being variables and rows being their observations. This matrix, shown in Table 1, together with the binary vector indicating which observations had been identified as SR patents, served as a training input for the machine learning approach.

Machine learning for classification analyses

Statistical classification using machine learning algorithms has long been implemented for the purpose of solving various problems and tasks such as computer vision, drug discovery or handwrite and speech recognition. Numerous different methods were developed and new ones still appear. However, there has been no one, at least to our knowledge, using statistical classifiers on the basis of a primal lexical query for the purpose of detecting an emerging technology. We considered a number of alternatives (Kotsiantis 2007) to the

[16] We also tried to incorporate another step (6), which added IPC dummy variables to indicate class belongings. These additional attributes were later abandoned by the following feature selection process, which suggests that these IPC class belongings are not significant for the categorization at hand.

Table 1 Structure of patent word and n-gram frequency matrix with binary decisions as input for machine training

Patent	Attribute vectors x									Binary decision y
	$word_{w1}$	$word_{w1}$	…	$bigram_{b1}$	$bigram_{b2}$	…	$trigram_{t1}$	$trigram_{t2}$	…	
1	$freq_{1\|w1}$	$freq_{1\|w2}$	…							1
2	$freq_{2\|w1}$	…								−1
…	…									…
205	$freq_{205\|w1}$	…			$freq_{205\|b2}$	…				−1
206	$freq_{206\|w1}$	…			…					1
…	….									…
228	$freq_{228\|w1}$	…						$freq_{228\|t2}$	…	−1
xxx	$freq_{xxx\|w1}$	…						…		?
…	…									

The rows 1 to 205 indicate an example of a subsample on which the machine is trained. The rows 206 to 228 then act as an example of a subset of data which is used for testing the fitness of the classification process. The last rows from xxx onwards at the bottom refer to new data, on which the SVM is able to decide based on the previous training.

aforementioned SVM, such as *k*-Nearest Neighbor, Neural Networks, and Genetic Algorithms before starting with our particular algorithm. According to the no-free-lunch theorem (Wolpert and Macready 1997), there is no general superior machine learning method and every problem has to be tackled individually depending on its properties. We therefore assessed the aforementioned algorithms according to run-time performance, sensitivity to irrelevant or redundant features, and ability to overcome local maximums. In a nutshell, SVM proved to be the most suitable algorithm and this decision was in line with computer science experts' opinions from robotics groups at the Karlsruhe Institute of Technology (KIT).

Support vector classification

The method of support vectors was introduced in the middle of the 1960s (Guyon et al. 1993; Cortes and Vapnik 1995). The original approach together with its various extensions is now one of the most acknowledged and recommended tools among modern machine learning algorithms. In the following, we briefly describe its core concept and discuss some advantages that are found to be relevant to the problem at hand. Simply put, the core idea of the method is to create a unique discrimination profile (represented by a linear function) between samples from (usually two[17]) different classes. The result is a line—or more generally a hyperplane—which is constructed in such a way that the distance between two parallel hyperplanes touching nearest samples becomes as large as possible. In this way, the method is trying to minimize false classification decisions.

The "touching" data points are termed support vectors. In fact, the resulting separation plane is shaped only by these constraining (=supporting) points. Below, we provide the

[17] There exist some multiclass SVM approaches. See Duan and Keerthi (2005) for a review.

mathematical notation of a support vector machine following Hsu et al. (2010), an article which is a comprehensive introduction to the method for purposes such as ours. Formally defined, we have a training set (x_i, y_i) of $i = 1,\ldots,l$ sample points (here: our patents), where every $x_i \in R^n$ is an attribute vector (consisting of our normalized word and n-gram frequencies) and $y_i \in \{-1, 1\}^l$ is a decision for that specific data point which thus defines its class. The SVM then yields the solution to the following optimization problem (see as well Boser et al. 1992; Guyon et al. 1993):

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

$$s.t.\ y_i\left(w^T \Phi(x_i) + b\right) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

in which W is the normal vector between the separating hyperplane and the parallel planes spanned by the support vectors. The mapping Φ is related to so-called Kernel functions, such that $K\left(x_i, x_j\right) \equiv \Phi(x_i)^T \Phi\left(x_j\right)$. For problems in which the data under consideration are not linearly separable, Φ maps the training attributes into a higher dimensional space where a hyperplane may be found. Table 2 summarizes common Kernel functions and their respective parameters γ, r, and d (Burges 1998; Ali and Smith-Miles 2006; Pedregosa et al. 2011; Manning et al. 2008).[18]

The above version of the classification procedure also incorporates the so-called *Soft-Margin* method (Cortes and Vapnik 1995) that allows for mislabeled training sample points. The approach introduces ξ_i as nonnegative slack variables which measure the extent of incorrectly classified items in the training set. $\sum_{i=1}^{l} \xi_i$ is thus a penalty term, and C a penalty parameter, on which we will comment later.

Training algorithm, classification, and evaluation

Figure 1 depicts the flow chart of our algorithm. First, we preprocessed the data in order to eliminate irrelevant features and to obtain a final dataset of feature vectors. When we turn to the result section, the necessity of this preprocessing becomes clearer. In a second step, we started the SVM training process.

It comprised three iterative steps that are found in almost any machine learning approach: Training of the model, its evaluation, and optimization. We realized all these steps for our SVM using the python programming language and its tool python scikit-learn for machine learning (Pedregosa et al. 2011).[19] Finally, the classifier with the best model fit was applied to some test data.

[18] Since there is no possibility to determine in advance which Kernel function should be used, the choice of the depicted functions was mostly motivated by their popularity in classifiers and availability within the software package used.

[19] We do not discuss the exact implementation of the support vector machine algorithm in the python scikit-learn tool. All necessary materials can be found in open access libraries following the reference provided above.

Table 2 Kernel functions used for the SVM

Kernel function	Formula
Polynomial	$(\gamma\langle x, x'\rangle + r)^d$
Radial basis function (rbf)	$\exp(-\gamma\lvert x - x'\rvert^2)$
Sigmoid	$\tanh(\langle x, x'\rangle + r)$

In order to avoid overfitting problems, we applied k-fold cross-validation during our training step. The algorithm, first, randomly splits the training dataset X into training and test parts. Second, it fits the model based on the training dataset leaving out the test data. During the training process, the data are again split into k parts. The algorithm then trains the model on k-1 parts and validates on the k-th part. The training is performed several times so that every part serves as a validation dataset. The number of training repetitions is reflected by a cross-validation parameter and can be specified. Thus, it is subject to variation during the overall fitting of the model itself. Figure 2 illustrates the k-fold cross-validation process.

The evaluation of our model is based on the criteria of precision and recall. The former criterion measures the ability of a classifier not to label objects as positive that should have been labeled negative. Formally, precision is the total number of true positives (tp) divided by the sum of all positives including false-positive errors (fp).

$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}$$

The latter criterion, recall, measures the ability of a classifier to find all positives or, again more formally, the number of true positives divided by a sum of true positives and false-negative errors (fn).

$$\text{Recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}$$

On the one hand, a model with a good recall but bad precision will find all positive samples—but will have some of them being actually negative. On the other hand, a model with bad recall but high precision will not have false-positive objects, however it will miss some of the true positives.

In order to balance these two measures, we used a so-called $f1$ score that can be seen as their weighted average:

$$f1 = 2 \times \frac{\text{Recall} \times \text{precision}}{\text{Recall} + \text{precision}}$$

To optimize our classifier, we calibrated it to have the highest possible $f1$ score. Tuning of the model was done by varying the cross-validation parameter, the kernel functions, and their respective parameters.

Results

The sample used in the machine learning process consisted of 228 patents with valid expert decisions. It contained 98 SR patents and 130 IR patents according to our expert group's validation. As a result of the transformation of unstructured patent text into structured data,

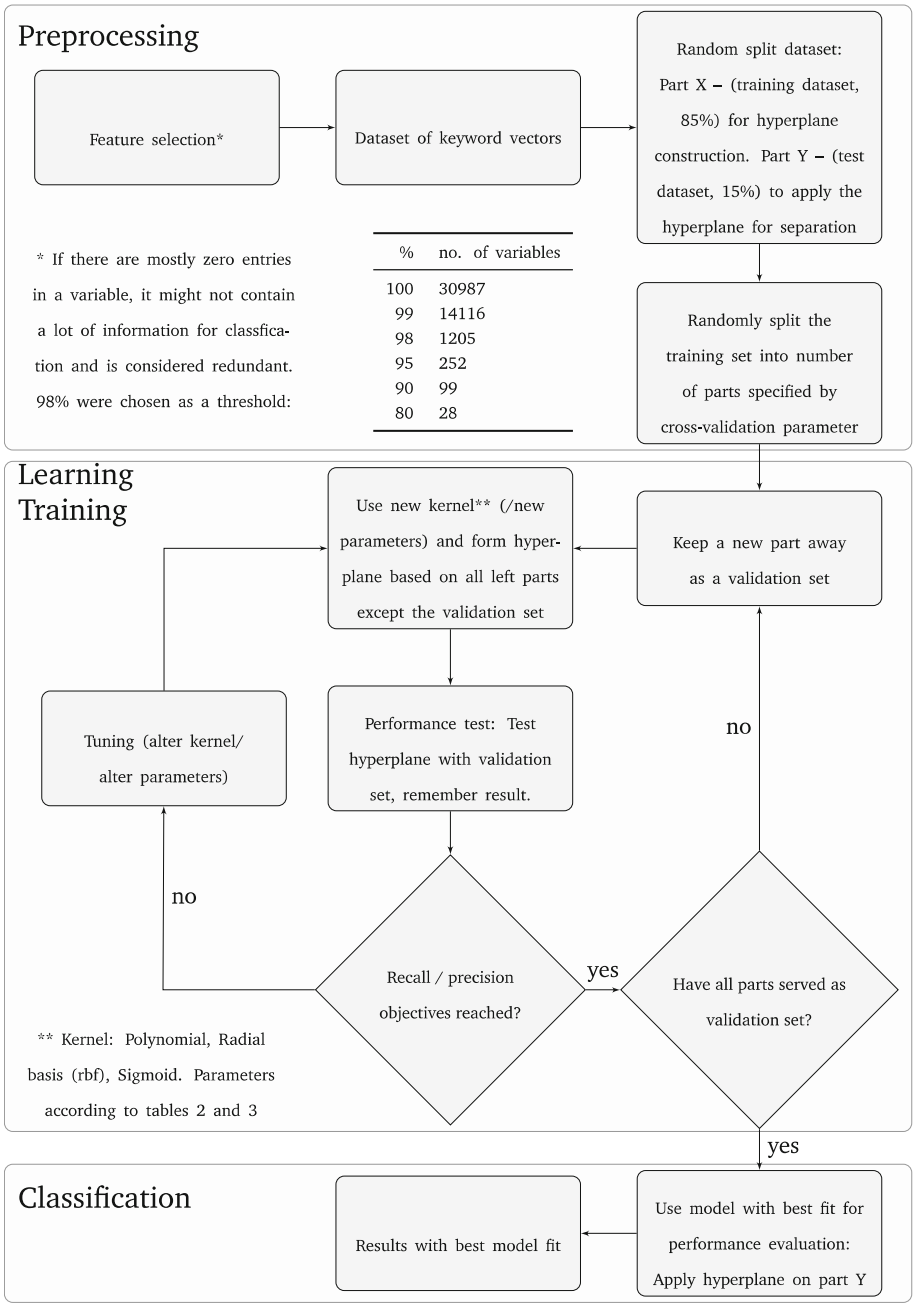

Fig. 1 Flow chart of the machine discrimination algorithm with preprocessing, support vector training, and final classification

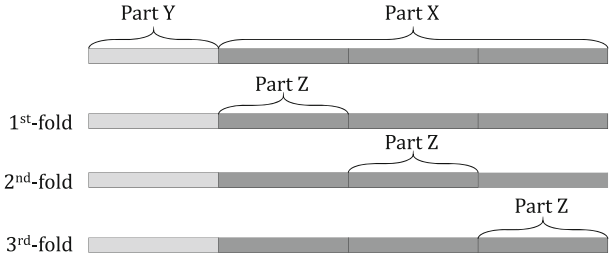

Fig. 2 *k*-fold cross-validation process

we observed 30,987 different features (or variables) within these patents, which included keywords, bigrams, and trigrams.[20] The resulting matrix $(228 \times 30,987)$ had to be pre-processed before serving as an input for the SVM, due to the fact that the majority of the variables contained zero entries. This means that only a small number of keywords and n-grams are shared by a majority of the patents. At first glance, this information could appear confusing. The explanation lies in the variety of SR applications: Descriptions of significantly different service robots with very unlike applications contain a huge number of dissimilar keywords and keyword combinations. Most of these are uniquely used in their specific contexts and thus appear with a very low frequency. Figure 3 illustrates this fact by showing typical relative appearances of normalized frequencies of two randomly chosen variables.

Thus, some variables contained too little information and introduced noise instead. Consequently, these insignificant features had to be excluded from the data for the purpose of improving the SVM performance. For example, if a keyword (or n-gram) appeared in only one patent, this variable would not have helped in solving the problem of classification. Our feature selection process served to exclude such a redundant feature. We implemented a threshold that at least 2% of the entries of a variable in each class (SR vs. IR) should have nonzero entries. The table in the flowchart (Fig. 1) shows the dependency between the number of variables and different thresholds. With this selection process, the resulting matrix was reduced to 1206 variables for our 228 observations/patents. We provide these variables/terms in Tables 10, 11, 12, 13, 14, 15 and 16 in the Appendix. Finally, all variable frequencies were scaled to the interval [0,1] such that a second normalization process set the maximum frequency in the sample to 1.

Figure 4 shows normalized frequencies of arbitrarily chosen attributes (keywords) in a two-dimensional space for all patents in the sample. The shape of the dots indicates the patent's expert classification as SR (square) and IR (diamond). Every subfigure thus represents one slice of the multi-dimensional space that the SVM tries to separate into two distinct zones with SR and IR patents. Figure 4 provides a graphical representation of the complexity of identifying apt discrimination lines at all—not to mention the determination of a multi-dimensional plane through the complete hypercube—without the utilization of machine learning algorithms.

[20] We even included IPC classes at an early stage of development, but did not find any of these classifications to become part of the support vectors. They turned out to be irrelevant to the discrimination procedure and were thus removed during the feature selection process.

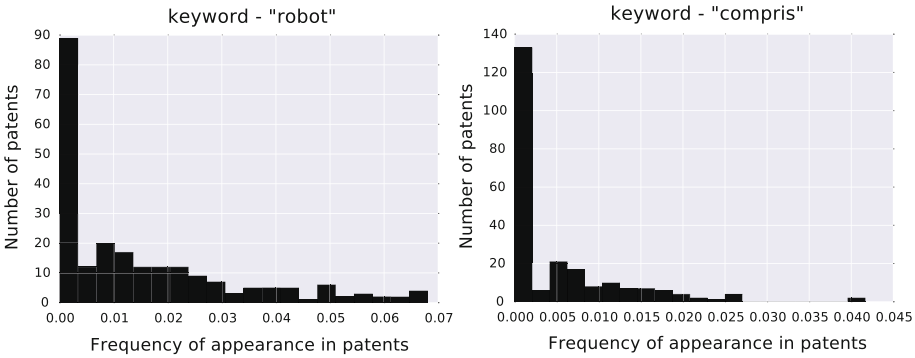

Fig. 3 Two representative frequency histograms of keywords showing that the majority of patents do not share the same terms

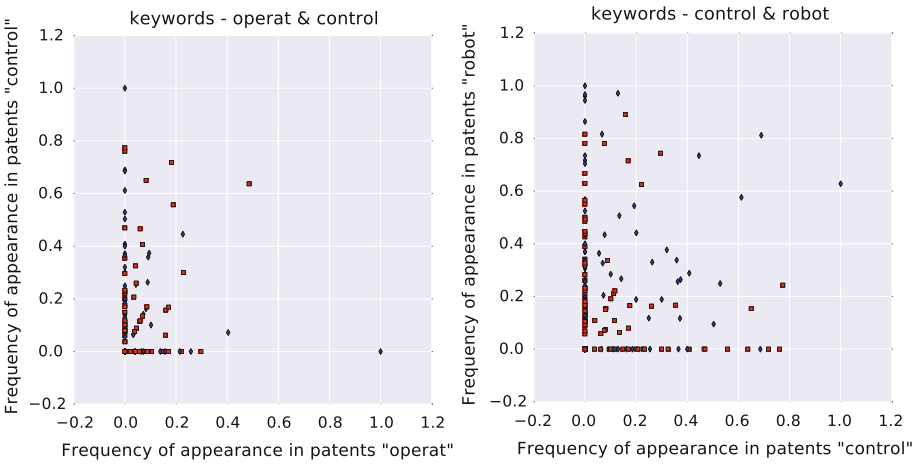

Fig. 4 Two representative pairs of keywords and their frequencies of appearance in IR (*diamond*) and SR (*square*) patents as classified by expert decisions. Note that appearance frequencies are normalized by the patent text length

SVM-specific outcomes

In order to eliminate negative influences of the unbalanced dataset, we introduced weights in our SVM proportionate to SR and IR classes. Following the cross-validation procedure, the support vector machine was fit on to an 85% of the original dataset. The remaining 15% were kept for testing purposes. The split was random, and its ratio is an arbitrary choice of authors.

The cross-validation parameter was set to 3 and 4 determining the amount of random splits of a training dataset into a training and evaluation set. Another parameter that was varied while searching for a better model was the so-called C parameter. The following citation nicely explains the main properties of this penalty parameter: “In the support vector network algorithm, one can control the trade-off between complexity of decision rule and frequency of error by changing the parameter C” (Cortes and Vapnik 1995, p. 286).

Table 3 Model tuning parameters and respective values

Parameter	Varied values	Chosen values
Cross-validation (cv)	$3, 4$	3
Complexity (C)	$10, \ldots, 1000$	10
γ of rbf kernel	$10^{-6}, \ldots, 10^{-2}$	0.005
γ of polynomial kernel	$10^{-6}, \ldots, 10^{-2}$	Not chosen
d of polynomial kernel	$1, 2, 3$	Not chosen
r of polynomial kernel	$1, 2, 3$	Not chosen
γ of sigmoid kernel	$10^{-6}, \ldots, 10^{-2}$	Not chosen
r of sigmoid kernel	$1, 2, 3$	Not chosen

Table 4 Classification report

	Precision (%)	Recall (%)	$f1$ score (%)	No. of patents in test set
SR	75	94	83	16
IR	93	74	82	19
Avg./total	85	83	83	35

Finally, the three different kernel functions from Table 2 were considered. In particular, the first was a polynomial function and its γ, degree, and r coefficient. The second was a radial basis function (rbf) and its γ constant. The third was a sigmoid function and its γ and r constants. Table 3 presents all kernel parameters and their values that were considered to find the best-performing classifier—as well as all eventually chosen values.

Exhaustive simulations with all possible combinations of the above-mentioned parameters yielded the best f1 score of the model. Our final model showed an 85% precision and 83% recall. It contained a radial basis function kernel with γ equal to 0.005 and C equal to 10. The training set was randomly split into 3 equal parts for cross validation. The resulting discrimination plane between the two classes of patents was constructed using 192 support vectors, meaning that only these sample observations were significant for classification. Table 4 presents a classification report after classifying the test set of our sample.

Scope for improvement

There is some scope for an even more precise technology identification. First, there is still room for increasing the performance of the SVM method, namely regarding the kernel functions. Although there have not been any successful attempts to introduce automatic kernel selection algorithms yet (Ali and Smith-Miles 2006), it is probably possible to find a better function for our problem at hand. Second, the support vector machine can be seen as a first-tier machine classifier that we just started with. Other methods like genetic algorithms, neural networks or boosting as well as their combinations could be applied in additional steps. Finally, applying principal component analyses (PCA) to our matrix of variables could provide some insights about a similar behavior of different key words in patents, which means that they could be grouped and analyzed together. Moreover,

applying PCA in SVM, we could say whether these groups of variables are significant in identifying an emerging technology—like service robotics in our case.

Conclusion

In this paper, we proposed a novel methodology for detecting early developments of an emerging technology in patent data. Our method uses a support vector machine algorithm on the example of robotics patents. The resulting model is able to find 83% of service robotics patents and classify them correctly with a probability of 85%.

There are several advantages of our method regarding technology classification tasks, which we will discuss along the criteria of Mogoutov and Kahane (2007) that we mentioned above: First, experts do not choose which terms should be added to or excluded from the primal search, hence the typical lexical bias towards preferred subfields is limited. Speaking of lexical versus citationist approaches, our method also avoids a major drawback of citational methods which circle around a core dataset and rely on future works explicitly referring to this prior art: Since citations in patents are generally rare,[21] for young emerging technologies in particular, the citation lag reduces the expected number of citations for any given document to a negligible amount.

Second, the procedure offers strong portability, such that it can easily be applied to scientific publications—taken for instance from Web of Science. Moreover, our step-by-step classification method can basically be applied to any emerging technology—not only to those arising as an initially small subset consisting of niche applications like SR emerging out of robotics. Nanotechnology, which in this respect is again a meaningful instance, would have been hard to detach from some well-defined mother technology. In fact, it became an umbrella term[22] for technological developments from various directions that had solely in common to work on a sufficiently small scale and to make intentionally use of the phenomena that arise on this scale. Nanotechnology thus consolidated endeavors from physics, chemistry, material technologies, and biology and had a converging character. The same is true for Industry 4.0, a term coined in Germany around 2010 that describes a fourth industrial revolution through automation and information networks in manufacturing technologies—or, in other terms, the Internet of Things in production. Within modular structured production environments, so-called cyber-physical systems such as robotic systems, digitally cross-linked machines, and even the 'intelligent' products themselves communicate with each other in real time, allowing both decentralized sequence control and centralized optimization regarding complex interdependencies throughout the complete value chain. This makes Industry 4.0 a superordinate concept describing digitally cross-linked production systems and thus enveloping various heterogeneous subtechnologies that are hardly classifiable. One of our future tasks will thus comprise the application of our method to historical nanotechnological patent sets as well as to Industry 4.0 technologies in order to demonstrate the general applicability and robustness of our method.

[21] Within PATSTAT, for instance, more than 90% of the listed patent applications are followed by less than three forward citations, 74% do not show any at all.

[22] SR may be seen as such an umbrella term as well—or as a system of technologies, i.e., combining many technologies in the way described by Arthur (2009): In that sense, SR is becoming more diverse and complex with evolving purposes organized to meet human needs.

Third, our algorithm approach shows high adaptability. Due to its learning nature it is able to produce valid outcomes although the technology under consideration is constantly evolving. Fourth and of capital importance, the proposed method performs well in terms of recall and precision scores, proving sufficient extent and relevance of the obtained data.

Acknowledgements We are thankful to the High Performance Humanoid Technologies (H2T) group from the Institute for Anthropomatics and Robotics at Karlsruhe Institute of Technology in Germany, in particular to Prof. Dr. Tamim Asfour and Prof. Dr. Gabriel Lopes from Delft Center for Systems and Control/Robotics Institute at TU Delft in the Netherlands for their support and advices. Moreover, we wish to thank the participants in the 15th EBES Conference in Lisbon, 6th annual S.NET meeting in Karlsruhe as well as 5th Global TechMining Conference in Atlanta for their valuable comments and suggestions that have led to the improvement of this article. This work is supported by the project "Value Creation & Innovation Processes in and beyond Technology" of the Karlsruhe School of Services.

Appendix

See Tables 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 and 16.

Table 5 Important robot definitions according to ISO 8373:2012

	Definition
Robot	Actuated mechanism programmable in two or more axes with a degree of autonomy, moving within its environment, to perform intended tasks *Note 1 to entry*: A robot includes the control system and interface of the control system *Note 2 to entry*: The classification of robot into industrial robot or service robot is done according to its intended application
Autonomy	Ability to perform intended tasks based on current state and sensing, without human intervention
Control system	Set of logic control and power functions which allows monitoring and control of the mechanical structure of the robot and communication with the environment (equipment and users)
Robotic device	Actuated mechanism fulfilling the characteristics of an industrial robot or a service robot, but lacking either the number of programmable axes or the degree of autonomy

Table 6 SR application examples of personal/domestic use according to IFR

	Definition
Robots for domestic tasks	Robot butler, companion, assistants, humanoids
	Vacuuming, floor cleaning
	Lawn mowing
	Pool cleaning
	Window cleaning
Entertainment robots and toy robots	Robot rides
	Pool cleaning
	Education and training
Handicap assistance and robotized wheelchairs	Personal rehabilitation
	Other assistance functions
Personal transportation	
Home security and surveillance	

Table 7 SR application examples of professional/commercial use according to (IFR, 2014)

	Applications
Field robotics	Agriculture
	Milking robots
	Forestry
	Mining systems
	Space robots
Professional cleaning	Floor cleaning
	Window and wall cleaning
	Tank, tube and pipe cleaning
	Hull cleaning education and training
Inspection and maintenance systems	Facilities, plants
	Tanks, tubes and pipes, and sewers
	Other inspection and maintenance systems
Construction and demolition	Nuclear demolition and dismantling
	Other demolition systems
	Construction support and maintenance
	Construction
Logistic systems	Courier/mail systems
	Factory logistics
	Cargo handling, outdoor logistics
	Other logistics
Medical robotics	Diagnostic systems
	Robot-assisted surgery or therapy
	Rehabilitation systems
	Other medical robots
Defense, rescue and security applications	Demining robots
	Fire- and bomb-fighting robots
	Surveillance/security robots
	Unmanned aerial and ground-based vehicles
Underwater systems	Search and rescue applications
	Other
Mobile platforms in general use	Wide variety of applications
Robot arms in general use	Wide variety of applications
Public relation robots	Hotel and restaurant robots
	Mobile guidance, information robots
	Robots in marketing
Special purpose	Refueling robots
Customized robots	Customized applications for consumers
Humanoids	Various applications

Table 8 Exemplary extract of robot patents under consideration with titles, publication numbers (given by the patent authority issuing the patent), filing dates (on which the application was received), and expert classification decisions

Title	Publication no.	Filing date	SR y/n?
Remote control manipulator	968525	1962-06-25	n (−1)
Folded robot	2061119	1979-10-24	n (−1)
In vivo accessories for minimally invasive robotic surgery	2002042620	2001-11-06	y (1)
Apparatus and method for nondestructive inspection of large structures	6907799	2001-11-13	y (1)
Surgical instrument	2002128661	2001-11-16	y (1)
Robotic vacuum cleaner	2003060928	2001-12-04	y (1)
A cleaning device	1230844	2002-01-21	n (−1)
Climbing robot for movement on smooth surfaces e.g., automatic cleaning of horizontal/vertical surfaces, has chassis with crawler drive suspended and mounted turnably about vertical axis to detect obstacles and prevent lifting-off	10212964	2002-03-22	y (1)
Single-cell operation supporting robot	2004015055	2002-08-08	y (1)
Underwater cleaning robot	2007105303	2006-03-14	y (1)
Position determination for medical devices with redundant position measurement and weighting to priorities measurements	1854425	2006-05-11	y (1)
Mobile robot and method of controlling the same	2007135736	2006-05-24	y (1)
Customizable robotic system	2012061932	2011-11-14	y (1)
Positioning apparatus for biomedical use	2012075571	2011-12-06	n (−1)
Apparatus and method of controlling operation of cleaner	2012086983	2011-12-19	n (−1)

Table 9 A fragment of a modular SQL Boolean term search approach for PATSTAT, defined through specific word construction for IFR application field CLEANING SR, augmented by IPC class codes

```
…
(
        (
        SUBSTRING(IPC.ipc_class_symbol,1,5)='"B08B'    OR
        SUBSTRING(IPC.ipc_class_symbol,1,5)='"E01H'    OR
        IPC.ipc_class_symbol LIKE '%B60S 1%            OR
        IPC.ipc_class_symbol LIKE '%B60S 3%'
        )
OR(
        TTL.appln_title LIKE '%robot%'                 AND(
                                                               TTL.appln_title LIKE '%suction cup%'
                                                               TTL.appln_title LIKE '%safety analy%'
                                                               TTL.appln_title LIKE '%vertical wall%'
                                                               TTL.appln_title LIKE '%dry adhesive%'
                                                               TTL.appln_title LIKE '%clean%'          AND NOT (…
```

AST refers to table containing abstracts, TTL refers to table containing titles, IPC refers to table containing IPC classes

Table 10 List of the 1206 variables used in the SVM for classification: Part 1/4 of the 726 unigrams

1a	arrang	cardiac	confirm
abl	arrangement	carri	connect
abnormal	arriv	carriag	connection
accelerat	articulat	carrier	consequent
access	assembl	caus	consist
accommodat	assist	cell	constitut
accord	associat	center	construct
accordanc	attach	centr	construction
accurat	attachabl	central	contact
achiev	attachment	chang	contain
acquir	auto	characteris	container
act	automat	characteristic	continuous
action	automatic	characteriz	control
activ	autonomous	charg	controller
actual	auxiliari	chassi	convention
actuat	avoid	check	convert
adapt	axe	circuit	conveyor
adapter	axi	claim	coordinat

Table 10 continued

addition	axial	clamp	correspond
adhesiv	backlash	clean	cost
adjacent	balanc	cleaner	coupl
adjust	barrier	climb	cover
adjustabl	base	clip	creat
adjustment	basi	close	crop
advanc	beam	coat	current
advantag	bear	code	customizabl
agricultural	behavior	collect	cut
aid	bend	collision	damag
aim	bicycl	column	data
air	bipedal	combin	decision
algorithm	blade	combinat	defin
allow	block	comfortabl	degre
amount	board	command	deliver
analysi	bodi	common	deliveri
analyz	bore	communic	deploy
angl	bottom	compact	depress
angular	box	compar	describ
animal	brush	compartment	design
annular	build	complementari	desir
apertur	built	complet	detachabl
apparatus	button	component	detect
appearanc	cabl	compos	detection
appli	calculat	compris	detector
applianc	camera	computer	determin
applic	capabl	condition	determinat
appropriat	capillari	configur	deviat
architectur	captur	configurat	devic
arm	car	confin	diagnosi

Table 11 List of the 1206 variables used in the SVM for classification: Part 2/4 of the 726 unigrams

differenc	endoscopic	form	inspection
difficult	energi	frame	instal
digital	engag	free	installat
dimension	enhanc	freedom	instruction
dimensional	ensur	frequenc	instrument
dip	enter	front	integrat
direct	entir	function	interaction
direction	environment	gear	interconnect
discharg	environmental	generat	interfac
disclos	equip	glove	interior
disconnect	equipment	grasp	internal

Table 11 continued

dispens	error	grip	invasiv
displac	especial	gripper	invention
displaceabl	essential	groov	involv
displacement	etc.	ground	item
display	exampl	guid	jet
dispos	exchang	guidanc	join
distal	exhaust	hand	joint
distanc	exist	handl	knee
dock	expensiv	har	laser
door	extend	head	latter
doubl	extension	heat	lawn
draw	external	held	layer
drill	extract	help	leg
drive	extraction	hip	length
driven	extrem	hold	lever
dust	facilitat	holder	lift
dynamic	faciliti	horizontal	light
earth	factor	hose	limb
easili	fasten	hous	limit
edg	featur	human	line
effect	feedback	hydraulic	linear
effectiv	field	identifi	link
effector	fig	imag	liquid
efficienc	figur	implement	load
elastic	fill	improv	local
electric	filter	improvement	locat
electronic	finger	includ	lock
element	fit	incorporat	locomotion
elongat	fix	increas	log
embodiment	flang	independent	longitudinal
emit	flat	individual	loop
emitter	flexibl	industrial	low
employ	floor	informat	lower
employment	flow	inner	machin
enabl	fluid	input	magnetic
enclos	forc	insert	main
endoscop	foreign	insertion	maintain

Table 12 List of the 1206 variables used in the SVM for classification: Part 3/4 of the 726 unigrams

make	obtain	portion	referenc
manipulat	oper	position	region
manner	operabl	possibl	register
manoeuvr	operat	power	relat
manual	oppos	pre	relationship
manufactur	optic	precis	relativ
map	option	predefin	releas
marker	orient	predetermin	reliabl
master	orientat	preferabl	remot
material	orthogonal	preparat	remov
mean	outer	press	removal
measur	output	pressur	replac
measurement	overal	prevent	requir
mechanic	pair	procedur	resolution
mechanism	pallet	process	respect
medic	panel	processor	respectiv
medicin	parallel	produc	result
medium	part	product	retain
memori	partial	production	return
method	particular	program	rigid
micro	pass	project	ring
militari	path	propos	risk
milk	patient	propulsion	robot
mine	pattern	protectiv	robotic
minimal	payload	provid	rock
mobil	perform	proximal	rod
modal	performanc	purpos	roll
mode	period	quantiti	roller
model	peripheral	rack	rotari
modul	permit	radar	rotat
monitor	perpendicular	radial	rotatabl
motion	photograph	radio	rough
motor	pick	rail	run
mount	piec	rais	safeti
movabl	pipe	rang	sampl
move	pivot	rapid	save
movement	pivotabl	reach	scale
mow	place	reaction	screen
mower	plan	real	seal
mri	plane	realiti	section
multi	plant	realiz	sector
multipl	plastic	rear	secur
navigat	plate	receiv	select
network	platform	receiver	send
normal	play	reciprocat	sens

Table 12 continued

nozzl	plural	recognition	sensor
object	pneumatic	record	sent
obstacl	port	reduc	separat

Table 13 List of the 1206 variables used in the SVM for classification: Part 4/4 of the 726 unigrams

sequenc	substantial	transmission	wire
seri	substrat	transmit	wireless
serv	subsystem	transmitter	workpiec
servo	suction	transport	worn
set	suitabl	transportat	wrist
shaft	suppli	transvers	zone
shape	support	travel	
shield	surfac	treat	
ship	surgeon	treatment	
short	surgeri	tube	
signal	surgic	type	
significant	surround	typic	
simpl	sutur	ultrasonic	
simulat	switch	underwater	
simultaneous	system	uneven	
singl	take	unit	
site	tank	universal	
situat	target	unload	
size	task	upper	
skin	techniqu	use	
slave	telepresenc	user	
sleev	telescopic	utiliz	
smooth	terminal	vacuum	
sourc	terrain	valu	
sow	test	variabl	
space	therebi	varieti	
spatial	therefrom	vehicl	
special	thereof	velociti	
specifi	thereon	vertic	
specific	thereto	vessel	
speed	third	video	
spiral	tight	view	
spray	tilt	virtual	
spring	time	visual	
stabiliti	tip	volum	
stabiliz	tissu	walk	
stabl	tool	wall	

Table 13 continued

stage	tooth	wast
station	top	water
stationari	torqu	weed
steer	torso	weight
step	touch	weld
stop	toy	wheel
storag	track	wherebi
store	train	wherein
structur	trajectori	wide
subject	transfer	winch
subsequent	translat	window

Table 14 List of the 1206 variables used in the SVM for classification: Part 1/2 of the 370 bigrams

1,2	button,effector	deviat,actual	imag,process
1,compris	capabl,control	devic,17	implement,method
1,computer	cardiac,procedur	devic,compris	includ,base
1,connect	chassi,frame	devic,control	includ,main
1,disclos	claim,includ	devic,determin	includ,pair
12,includ	clean,horizontal	devic,direct	includ,step
12,provid	clean,method	devic,includ	independent,claim
13,14	clean,operat	devic,main	industrial,robot
2,3	clean,robot	devic,position	informat,relat
2,compris	cleaner,compris	devic,provid	informat,sensor
2,move	cleaner,invention	devic,robot	informat,set
3,4	comfortabl,position	devic,system	inner,surfac
3,compris	component,provid	direction,drive	input,button
3,connect	compris,base	displacement,sensor	input,data
4,5	compris,bodi	distanc,measur	instrument,coupl
43,connect	compris,main	door,10	instrument,effector
5,arrang	compris,plural	drive,actuat	instrument,mount
5,provid	compris,robot	drive,devic	invasiv,cardiac
accord,invention	compris,robotic	drive,forc	invention,compris
actual,position	computer,program	drive,ground	invention,disclos
actuat,control	connect,clamp	drive,mechanism	invention,propos
addition,equipment	control,box	drive,system	invention,provid
adjust,position	control,cabl	drive,unit	invention,relat
adjustabl,surgeon	control,devic	drive,wheel	joint,provid
allow,surgeon	control,input	e,g	laser,emitter
angl,adjust	control,joint	effector,control	leg,joint
apparatus,compris	control,manipulat	effector,correspond	longitudinal,direction
apparatus,method	control,method	effector,handl	machin,tool
apparatus,perform	control,movement	effector,manipulat	main,bodi
arm,coupl	control,operat	effector,move	main,controller

Table 14 continued

arm,includ	control,panel	effector,movement	manipulat,arm
arm,instrument	control,provid	effector,perform	manipulat,hold
arm,join	control,resolution	element,5	master,handl
assembl,method	control,robot	endoscopic,imag	mean,14
automatic,clean	control,robotic	error,signal	mean,2
automatic,control	control,system	factor,adjustabl	mean,detect
automatic,robot	control,unit	front,bodi	mean,receiv
autonomous,move	controller,handl	front,rear	measur,devic
autonomous,robot	correspond,movement	front,robot	mechanism,rotat
axe,rotat	coupl,pair	guid,mean	method,apparatus
balanc,control	degre,freedom	hand,surgeon	method,autonomous
base,informat	deliveri,system	handl,controller	method,clean
base,station	depress,surgeon	handl,move	method,control
bodi,2	detect,obstacl	handl,scale	method,invention
bodi,robot	detect,position	har,1	method,provid
bodi,surgic	detection,mean	hold,sutur	method,system
button,allow	determin,position	horizontal,vertic	method,thereof
button,depress	determin,spatial	imag,data	method,use

Table 15 List of the 1206 variables used in the SVM for classification: Part 2/2 of the 370 bigrams

minimal,invasiv	position,coordinat	robot,pick	system,includ
mobil,robot	position,determinat	robot,position	system,method
mobil,robotic	position,devic	robot,realiz	system,mobil
motion,control	position,handl	robot,robot	system,perform
motion,controller	position,informat	robot,s	system,robot
motor,drive	position,robot	robot,system	system,use
motor,vehicl	position,robotic	robotic,arm	thereof,invention
mount,chassi	position,system	robotic,control	time,period
mount,robot	power,sourc	robotic,devic	tissu,robotic
move,button	predetermin,position	robotic,surgeri	travel,perform
move,comfortabl	predetermin,time	robotic,system	tube,apparatus
move,devic	procedur,system	rotari,brush	typic,movement
move,effector	produc,correspond	rotat,axe	uneven,terrain
move,floor	provid,mean	rotat,head	unit,arrang
move,robot	provid,platform	rotat,motor	unit,compris
move,surgeon	provid,robot	rotat,movement	unit,control
movement,effector	provid,surgic	rotat,shaft	unit,drive
movement,handl	purpos,robot	scale,effector	unit,generat
movement,movement	real,time	scale,factor	unit,provid
movement,perform	relat,automatic	seal,access	upper,lower
movement,robotic	relat,method	send,imag	use,robotic
movement,typic	relat,mobil	sensor,mount	use,surgic
navigat,system	relat,robot	servo,motor	user,operat

Table 15 continued

object,provid	remot,control	signal,receiv	vacuum,clean
operat,accord	remot,view	signal,robot	vacuum,cleaner
operat,clamp	resolution,effector	signal,transmitter	vehicl,bodi
operat,devic	robot,1	slave,robot	vertic,axi
operat,operat	robot,10	smooth,surfac	video,signal
operat,perform	robot,arm	sow,weed	walk,robot
operat,power	robot,arrang	surfac,clean	water,discharg
operat,rang	robot,automatic	surgeon,adjust	wheel,instal
operat,remot	robot,bodi	surgeon,control	wire,wireless
operat,robot	robot,capabl	surgeon,input	x,y
operat,unit	robot,clean	surgeon,produc	y,z
output,signal	robot,cleaner	surgeon,scale	
overal,structur	robot,communic	surgeri,surgic	
pair,master	robot,compris	surgic,instrument	
pair,robotic	robot,control	surgic,operat	
pair,surgic	robot,includ	surgic,procedur	
path,robot	robot,invention	surgic,robot	
patient,s	robot,main	surgic,site	
patient,treat	robot,method	surgic,system	
perform,clean	robot,mobil	surgic,tool	
perform,hand	robot,motion	sutur,tissu	
perform,minimal	robot,move	system,autonomous	
perform,surgic	robot,movement	system,compris	
position,base	robot,mower	system,control	
position,compris	robot,operat	system,devic	

Table 16 List of the 1206 variables used in the SVM for classification: All 110 trigrams

adjust,position,handl	invention,relat,automatic	surgeon,produc,correspond
adjustabl,surgeon,control	invention,relat,method	surgeon,scale,factor
allow,surgeon,adjust	invention,relat,mobil	surgic,instrument,coupl
apparatus,perform,minimal	manipulat,hold,sutur	surgic,instrument,mount
arm,coupl,pair	master,handl,controller	surgic,robot,compris
arm,instrument,effector	method,invention,relat	surgic,robot,system
button,allow,surgeon	method,thereof,invention	sutur,tissu,robotic
button,depress,surgeon	minimal,invasiv,cardiac	system,control,movement
button,effector,move	mobil,robot,invention	system,includ,pair
cardiac,procedur,system	mobil,robotic,devic	system,perform,minimal
clean,horizontal,vertic	mount,robot,arm	thereof,invention,disclos
clean,robot,1	move,button,depress	tissu,robotic,arm
cleaner,invention,relat	move,comfortabl,position	typic,movement,perform
compris,main,bodi	move,effector,handl	x,y,z
control,input,button	move,surgeon,produc	surgeon,input,button
control,method,thereof	movement,effector,control	surgeon,control,resolution

Table 16 continued

control,resolution,effector	movement,effector,movement	instrument,effector,manipulat
controller,handl,move	movement,handl,scale	invasiv,cardiac,procedur
correspond,movement,effector	movement,movement,effector	surgeon,adjust,position
correspond,movement,typic	movement,perform,hand	instrument,coupl,pair
coupl,pair,master	movement,typic,movement	
coupl,pair,robotic	pair,master,handl	
depress,surgeon,input	pair,robotic,arm	
devic,main,controller	pair,surgic,instrument	
devic,robot,arm	perform,clean,operat	
effector,control,input	perform,hand,surgeon	
effector,correspond,movement	perform,minimal,invasiv	
effector,handl,move	position,handl,move	
effector,manipulat,hold	position,robot,arm	
effector,move,button	procedur,system,includ	
effector,movement,handl	produc,correspond,movement	
effector,movement,movement	relat,automatic,robot	
factor,adjustabl,surgeon	resolution,effector,movement	
front,robot,arm	robot,arm,includ	
hand,surgeon,scale	robot,cleaner,compris	
handl,controller,handl	robot,cleaner,invention	
handl,move,comfortabl	robot,control,method	
handl,move,effector	robot,control,system	
handl,move,surgeon	robot,invention,relat	
handl,scale,effector	robot,system,method	
hold,sutur,tissu	robotic,arm,coupl	
includ,pair,surgic	robotic,arm,instrument	
independent,claim,includ	robotic,devic,compris	
input,button,allow	scale,effector,correspond	
input,button,effector	scale,factor,adjustabl	

References

Abraham, B., & Morita, S. (2001). Innovation assessment through patent analysis. *Technovation, 21,* 245–252.

Ali, S. & Smith-Miles, K. A. (2006). A meta-learning approach to automatic kernel selection for support vector machines, *Neurocomputing* ***70*** (123), 173–186. Neural networks selected papers from the 7th Brazilian Symposium on Neural Networks (SBRN 04).

Arora, S. K., Porter, A. L., Youtie, J., & Shapira, P. (2013). Capturing new developments in an emerging technology: An updated search strategy for identifying nanotechnology research outputs. *Scientometrics, 95,* 351–370.

Arora, S. K., Youtie, J., Carley, S., Porter, A. L., & Shapira, P. (2014). Measuring the development of a common scientific lexicon in nanotechnology. *Journal of Nanoparticle Research, 16*(2194), 1–11.

Arthur, B. (1999). Complexity and the economy. *Science, 284*(5411), 107–109.

Arthur, B. (2009). *The nature of technology—what it is and how it evolves*. New York: Free Press. (Reprint Ed. January 11, 2011).

Arthur, B., & Polak, W. (2006). The evolution of technology in a simple computer model. *Complexity, 11*(5), 23–31.

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics, 118*(4), 1279–1333.

Bassecoulard, E., Lelu, A., & Zitt, M. (2007). Mapping nanosciences by citation flows: a preliminary analysis. *Scientometrics, 70,* 859–880.

Boser, B., Guyon, I. & Vapnik, V. (Eds.). (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory—COLT'92* (p. 144).

Bresnahan, T. F. (2010). General purpose technologies. In B. Hall & N. Rosenberg (Eds.), *Handbook of economics of innovation* (2nd ed., pp. 763–791). Amsterdam: Elsevier.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2), 121–167.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Cozzens, S., Gatchair, S., Kang, J., Kim, K.-S., Lee, H. J., Ordónez, G., et al. (2010). Emerging technologies: Quantitative identification and measurement. *Technology Analysis & Strategic Management, 22*(3), 361–376.

Duan, K.-B., & Keerthi, S. (2005). Which is the best multiclass SVM method? An empirical study. In N. Oza, R. Polikar, J. Kittler, & F. Roli (Eds.), *Multiple classifier systems* (Vol. 3541, pp. 278–285)., Lecture notes in computer science Berlin: Springer.

Erdi, P., Makovi, K., Smomogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P., et al. (2013). Prediction of emerging technologies based on analysis of the US patent citation network. *Scientometrics, 95*, 225–242.

Fischer, M., Scherngell, T., & Jansenberger, E. (2009). Geographic localisation of knowledge spillovers: Evidence from high-tech patent citations in Europe. *Annals of Regional Science, 43,* 839–858.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerization? *Technological Forecasting and Social Change*, *114*, 254–280.

Frietsch, R. (2015). Collection and analysis of private R&D investment and patent data in different sectors, thematic areas and societal challenges, JRC/BRU/2014/J.6/0015/OC, Inception report, Deliverable 1.1: Methodological report, Karlsruhe Fraunhofer Institute for Systems and Innovations Research.

Fu, L. D., & Aliferis, C. F. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in biomedical literature. *Scientometrics, 85,* 257.

Garfield, E. (1967). Primordial concepts citation indexing and historio-bibliography. *Journal of Library History, 2,* 235–249.

Graetz, G., & Michaels, G. (2015). Robots at work. *Center for Economic Performance Discussion Paper.*

Griliches, Z. (1990). Patent statistics as economic indicators: A survey. *Journal of Economic Literature, 28,* 1661–1707.

Guyon, I., Boser, B., & Vapnik, V. (1993). *Automatic capacity tuning of very large VC-dimension classifiers, advances in neural information processing systems* (pp. 147–155). Burlington: Morgan Kaufmann.

Halaweh, M. (2013). Emerging technology: What is it? *Journal of Technology Management and Innovation, 8*(3), 108–115.

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of Economics, 36*(1), 16–38.

Hsu, C.-W., Chang, C.-C. & Lin, C.-J. (2010). A practical guide to support vector classification. Technical Report, Department of Computer Science and Information Engineering, National Taiwan University.

Jaffe, A., Trajtenberg, M., & Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics, 108*(3), 577–598.

Kenekayoro, P., Buckley, K., & Thelwall, M. (2014). Automatic classification of academic web page types. *Scientometrics, 101,* 1015.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica, 31,* 249–268.

Lee, W. H. (2008). How to identify emerging research fields using scientometrics: An example in the field of information security. *Scientometrics, 76*(3), 503–525.

Lee, P., Su, H., & Chan, T. (2010). Assessment of ontology-based knowledge network formation by vector-space model. *Scientometrics, 85,* 689.

Lee, S., Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation, 29,* 481–497.

Li, Y.-R., Wang, L.-H., & Hong, C.-F. (2009). Extracting the significant-rare keywords for patent analysis. *Expert Systems with Applications, 36,* 5200–5204.

Liu, S., & Shyu, J. (1997). Strategic planning for technology development with patent analysis. *International Journal of Technology Management, 13,* 661–680.

Manning, C., Raghavan, P. & Schütze, H. (2008). Introduction to Information retrieval. online, Accessed Oct 15 2014. URL: http://www-nlp.stanford.edu/IR-book/.

McKeown, K. et al. (2016). Predicting the impact of scientific concepts using full-text features. *Journal of the Association for Information Science and Technology*, *67*(11), 2684–2696.

Mogoutov, A., & Kahane, B. (2007). Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking. *Research Policy, 36,* 893–903.

Noyons, E., Buter, R., Raan, A., Schmoch, U., Heinze, T., S., H. & Rangnow, R. (2003). Mapping excellence in science and technology across Europe. Part 2: nanoscience and nanotechnology. Draft Report EC-PPN CT2002-0001 to the European Commission. Leiden University Centre for Science and Technology Studies/Karlsruhe Fraunhofer Institute for Systems and Innovations Research.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12,* 2825–2830.

Porter, A., Youtie, J., & Shapira, P. (2008). Nanotechnology publications and citations by leading countries and blocs. *Journal of Nanoparticle Research, 10,* 981–986.

Ruffaldi, E., Sani, E. & Bergamasco, M. (2010). Visualizing perspectives and trends in robotics based on patent mining In *IEEE International Conference on Robotics and Automation*, Anchorage, Alaska.

Schmoch, U. (2008). Concept of a technology classification for country comparisons. In *Final Report to the World Intellectual Property Organization (WIPO)*, Karlsruhe Fraunhofer Institute for Systems and Innovations Research.

Srinivasan, R. (2008). Sources, characteristics and effects of emerging technologies: Research opportunities in innovation. *Industrial Marketing Management, 37,* 633–640.

Stahl, B. (2011). *What does the future hold? A critical view of emerging information and communication technologies and their social consequences, vol. 356 of Researching the Future in Information Systems, IFIP advances in information and communication technology*. Berlin: Springer.

Thompson, P. (2006). Patent citations and the geography of knowledge spillovers: Evidence from inventor- and examiner-added citations. *The Review of Economics and Statistics, 88*(2), 383–388.

Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Information Processing and Management, 43,* 1216–1247.

Van de Velde, E., Debergh, P., Verbeek, A., Rammer, C., Cremers, K., Schliessler, P., et al. (2013). *Production and trade in KETs-based products: The EU position in global value chains and specialization patterns within the EU*. Brussels: European Commission, DG Enterprise.

Wolpert, D., & Macready, W. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67–82.

Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *Journal of High-Technology Management Research, 15,* 37–50.