Contents lists available at ScienceDirect

Technological Forecasting & Social Change

Use of a new patent text-mining and visualization method for identifying patenting patterns over time: Concept, method and test application

Helen Niemann, Martin G. Moehrle *, Jonas Frischkorn

University of Bremen, IPMI – Institute of Project Management and Innovation, P.O. Box 330 440, D-28344 Bremen, Germany

ARTICLE INFO

Article history:
Received 17 March 2015
Received in revised form 29 June 2016
Accepted 3 October 2016
Available online xxxx

Keywords:
Patent lanes
Patent analysis
Cluster analysis
Similarity measurement
Visualization
Citation networks
Technological trajectories

ABSTRACT

Understanding the evolution of a technological field in the course of time is a key task in technology analysis. Analysts in research institutions as well as in companies need to know which topics are relevant for the respective technological field, which are the emerging topics, which traditional topics have been deepened in the course of time and which have been abandoned. For this purpose we suggest a patent lane analysis. Patent lanes can be seen as the deployment of patent clusters in the course of time. We use a method based on semantic similarities to develop patent lanes. A case study focuses on the application of carbon fibers in bicycle technology; it is used to demonstrate our method, i.e. to establish patent lanes in this case and characterize them by multiple use of a Tf idf measure. Despite some limitations, patent lanes enable deep insights into the development of patent-friendly technological fields.

1. Introduction

A lot of corporate technology managers and scientists in research and political institutions seek to understand typical evolutionary patterns of technological fields. For instance, they might wish to know which topics are relevant for the respective technological field, which are the emerging topics, which traditional topics have been deepened in the course of time and which have been abandoned? With respect to many, though by no means all, technological fields, patent analyses may help to answer such questions. They have been used successfully and extensively in many cases, and comprise different techniques, such as co-classification analysis (see as examples Choi et al., 2007; Chang et al., 2009; Dereli and Durmusoglu, 2009) or citation analysis (Tseng et al., 2011; Frietsch, 2007; Kuusi and Meyer, 2007; de Souza Carvalho et al., 2009; Lee et al., 2009, 2012).

A multitude of techniques for patent analysis makes use of the so-called meta-data of patents. Meta-data are defined by patent laws like the U.S. Code Title 35 and comprise information on applicants, inventors, classifications (international patent classification [IPC], current patent classification [CPC], in some cases national classifications like the US patent classification [USPC]), application and granting dates, cited patents and other literature (Ernst, 2003; Lee et al., 2011). Valuable answers to some of the questions mentioned above can be obtained by such analyses; techniques like activity analysis, co-classification activity analysis and citation network analysis may provide answers regarding different technical aspects and the development of topics in the technological field over time. The answers to some questions are not quite perfect yet, and there is still potential for improvement. Especially exploiting the information contained in the full-text of patents (instead or in addition to meta-data) by means of text mining technologies, as suggested by Yoon and Kim (2011) as well as by Moehrle and Gerken (2012) and Gerken and Moehrle (2012), may provide researchers with deeper insights. Text mining offers the opportunity to establish semantic similarity measures between documents and in doing so provides an alternative or an addition to the well known citation analysis.

In this paper we concentrate on these text mining technologies and suggest so-called patent lanes which we define as the deployment of patent clusters over time. The idea of patent lanes is related to the timeline visualization of the development of technological clusters (Shibata et al., 2010), but uses disaggregated information instead of clusters. The paper is organized as follows: In the next section we briefly explain how to measure semantic similarities between patents, as this constitutes the foundation of our method. In order to underpin our methodical contribution, we compare semantic similarities and citations as basic elements that establish links between patents, and show their interrelation in analyses. As patent lanes may be configured in different ways, we discuss the most important design decisions. A case study which focuses on carbon fiber reinforcements and the utilization thereof in bicycle technology, serves to illustrate the use of patent lanes and the

* Corresponding author.
E-mail addresses: helen.niemann@innovation.uni-bremen.de (H. Niemann), martin.moehrle@innovation.uni-bremen (M.G. Moehrle), jonas.frischkorn@innovation.uni-bremen (J. Frischkorn).

http://dx.doi.org/10.1016/j.techfore.2016.10.004
0040-1625/© 2016 Elsevier Inc. All rights reserved.

interpretation of results. We compare our method with methods characterized by rolling clustering to identify criteria for the usefulness of its application. Some concluding remarks will highlight implications as well as limitations of our method.

2. Measurement of semantic similarities between patents

One basic feature of our method is the application of semantic similarities between patents (see Moehrle, 2010). There is one major idea behind this: We assume that similarities between the contents of patents are reflected by similarities in language, for instance by the use of similar terminology (e.g. specifically scientific terminology), explanations of similar application situations, or a focus on similar useful functions. There seems to be some evidence to support this assumption, as in a recent study Möller and Moehrle (2015) have shown that this type of background information may significantly supplement and improve traditional keyword-based patent searches.

In the available literature several methods of measuring semantic similarities can be found. Having generated a basic set of patents representing the technology under investigation by means of keyword or classification-based search, the related tasks may be summarized as a generic process in four steps (see Moehrle and Gerken, 2012, see Fig. 1), comprising (i) preliminary language processing, (ii) concept extraction and building, (iii) variable measurement, and (iv) similarity calculation.

Before semantic measurements can take place, the data should be cleaned, i.e. terms should be reduced to their word-stems, synonyms should be harmonized, and filters for non-discriminant terms should be applied.

There are different ways to extract and build concepts (in the sense of key terms) from patent documents. Yoon and Kim (2011) use subject-action-object structures (SAOs) for this purpose and make use of knowledge about the syntactical functions of the extracted concepts. Moehrle and Gerken (2012) apply n-grams to generate solitary and combined concepts and give advice on how to configure the extraction. In this paper we concentrate on the latter option.

After extracting semantic concepts one way or the other, different variables can be measured. Such variables may represent the size of a patent, the overall overlapping set, or an overlapping set measured from the perspective of a pair of patents (double single-sided, abbreviated DSS, see Moehrle, 2010).

Based on the established variables established, semantic similarities can be calculated. "Similarity is formally defined as an increasing function of commonality and decreasing function of differences among objects to be compared" (Jeong et al., 2008). Different formulas are available for this purpose (see Gower and Legendre, 1986), for instance the Jaccard index that relates an overlapping set of terms to the sum of patent related sets of terms, or the Inclusion index that relates an overlapping set of terms to the patent related set of terms of the smaller patent.

3. Similarities and differences in connections between pairs of patents based on citations and semantic similarities

Having introduced semantic similarities, we now compare semantic similarities and citation analysis to underpin our methodical contribution. For this purpose we focus on the characteristics of semantic similarities and citations; later on we show the use thereof in different analyses.

Patent citations are generally differentiated into forward and backward citations. "Forward citations are the number of citations received by a patent. Counting the forward citations of patents shows whether patented inventions are mentioned – either by examiners … or by applicants or their lawyers" (Rost, 2011). In contrast, "backward citations are made by a patent to a previously issued patent. Studies using backward citation information investigated spillovers … between technology classes … or regions" (Rost, 2011). Patent citations have been used since the 1990s for establishing the importance of patents (see the work by Jaffe et al., 1993) and more recently for analyzing knowledge flows based on complexity theory (see Sorenson et al., 1993).

Basically, both citation based and semantic similarity based approaches connect pairs of patents. In the following we will first focus on the connection between a pair of patents, and then briefly discuss superior network structures.

There are five major differences in the connections between pairs of patents based on citations or on semantic similarities. They differ (i) in the range of values of the connection, (ii) the establishment of the connection, (iii) the timely availability of the connection, (iv) the foundation of the connection, and (v) the localization of the connection in a patent's parts (see Table 1). Compared to citations, semantic similarities comprise a continuous range of values, they are caused by lawyers and inventors who formulate the patent's wording (which leads to a fuzziness of the approach that has to be taken into account), they are completely available on the issue date of a patent, the connection elements can be identified (as the set of shared terms), and the connection elements can be located in the parts of a patent such as the claims

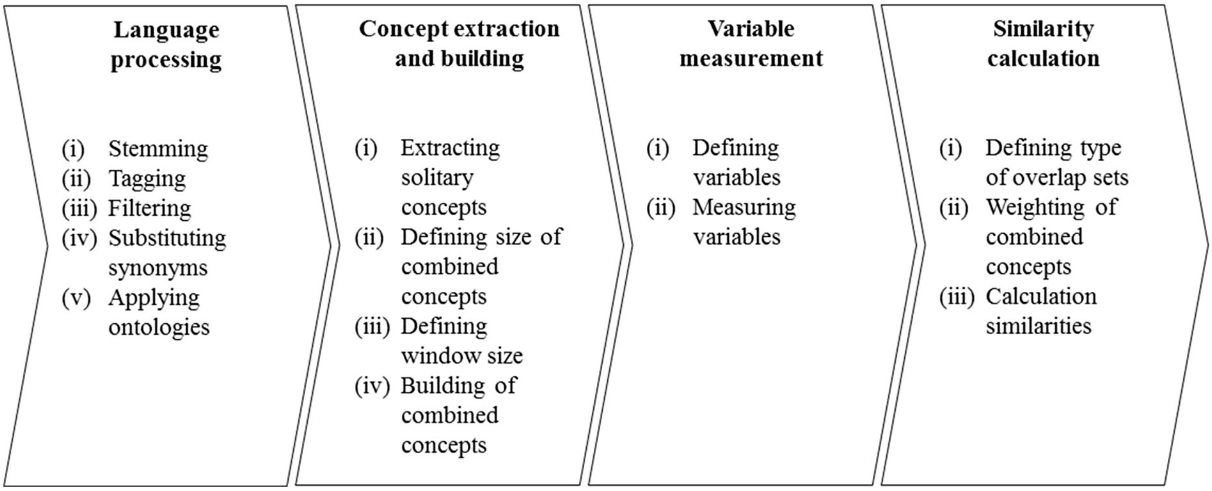

Language processing
(i) Stemming
(ii) Tagging
(iii) Filtering
(iv) Substituting synonyms
(v) Applying ontologies
Concept extraction and building
(i) Extracting solitary concepts
(ii) Defining size of combined concepts
(iii) Defining window size
(iv) Building of combined concepts
Variable measurement
(i) Defining variables
(ii) Measuring variables
Similarity calculation
(i) Defining type of overlap sets
(ii) Weighting of combined concepts
(iii) Calculation similarities

Fig. 1. Generic process for semantic similarity calculation.
Source: Moehrle and Gerken (2012), p. 807.

Table 1
Characteristics of connections based either on citations or semantic similarities. *Source: authors.*

Criterion	Citation connection	Semantic similarity connection
Value range of connection between a pair of patents	Dichotomous connection	Continuous connection between 0 and 1[*]
Establishment of connection	▪ Citation by inventor/applicant ▪ Citation of patent assessor	Measurement of semantic similarities, different options possible
Availability of connections	▪ Backward citations immediately with disclosure of patent ▪ Forward citations after a minimum of 18 month, no maximum[**,***]	Semantic similarities immediately with disclosure of patent
Foundation of connection	Not available	Semantic concepts that are responsible for similarities between patents may be extracted
Localization of connections in patent's parts	Not available	Different measurements available after separation of patent's parts

[*] In some constellations of similarity measurement the value may exceed 1, see Moehrle (2010).

[**] 70% of all patents are cited less often than three times; see Lee et al. (2009).

[***] Karvonen and Kässi (2013) use the so-called technology cycle time to analyze different industries. The technology cycle time is defined "as the median age of the patents cited on the front page of a patent document" (Kayal, 1999). For instance, Karvonen and Kässi (2013) find the technology cycle time for vertically integrated electronics to be 5.25 years, for paper printing 7.57 years, and for downstream electronics 4.89 years.

(even independent and dependent claims), the description, the abstract, or the title.

The relationship between using semantic similarities and citations as basic elements for establishing links between patents is twofold: Under specific conditions, citations may replace semantic similarities, and vice versa.

Citations may be used alternatively to generate a matrix of paired similarities between patents. Following the approach of co-citation analysis (see Lai and Wu, 2005), three steps have to be taken to obtain this type of matrix: (i) Similar to the semantic approach, a technological field has to be selected by means of a keyword or classification-based patent search. (ii) In contrast to the semantic approach, the resulting set of patents is divided into target patents (which cite other patents), basic patents (which are frequently cited) and others (which are considered irrelevant). (iii) co-citations are measured regarding basic patents only, and the linkage strength is calculated by application of a Jaccard formula. Based on the generated matrix of paired similarities regarding basic patents, a patent lane analysis may be executed (which would be based on citations).

Vice versa, semantic similarities may alternatively be used in approaches that are traditionally based on citations. For instance, the additional steps mentioned by Lai and Wu (2005) to gain insight into the topics related to a specific technological field might as well be executed on the basis of semantic similarities instead of citations, i.e. by using Pearson's correlation coefficient to harmonize the similarities, and applying a factor analysis to extract sets of patents which represent major topics in the technological field. Furthermore, semantic similarities may be seen as directed connections, assuming a knowledge flow from elder to younger patents (in analogy as it is implied with citations). Doing so, these connections between pairs of patents may be analyzed in a network approach (in analogy to Boyack and Klavans, 2010), looking for characteristics like centrality (in different forms) or density (Gilsing et al., 2008; Opsahl et al., 2010).

This would also offer opportunities to validate relevant publications by means of semantic similarities. For instance, von Wartburg et al. (2005) introduced multi-stage patent citation analysis to measure inventive progress. They suggested bibliographic couplings for multi-stages, distinguished between core and non-core patents based on network analyses, and used UCINET for calculating measures. More recently, Glänzel and Thijs (2012) developed a method to identify emerging topics in a field of science based on core documents that are generated by bibliographic coupling (the documents in question do not necessarily have to be patent documents; in their paper Glänzel and Thijs use scientific papers). Both could be validated by using semantic similarities instead of citations.

4. Concept and design decisions for patent lanes

Having analyzed some features of citations and their differences to semantic similarities, we will now introduce the concept of patent lanes. After outlining the basic idea and the related process, we will go into detail regarding the most important design decisions.

4.1. The concept of patent lanes

The basic idea behind patent lanes is to deploy a patent cluster over time in a specific way using semantic similarities between patents (see Niemann and Moehrle, 2013). In principle, the process of developing patent lanes comprises five steps (Fig. 2).

(i) A number of patents have to be defined as a basic set for the patent lanes. The selection may be achieved by keyword- or classification-based search.
(ii) Semantic similarities between all pairs of patents are measured to derive a similarity matrix. This may be done in different ways.
(iii) The oldest patents of the basic set form the starting set. By means of a number of cluster analyses, outliers and basic clusters are identified. Combined, these serve to form the starting lanes.
(iv) All other patents are treated in chronological order. Regarding each patent the semantic similarities with all older patents are considered and both the maximum value of the semantic similarity and the patent connected to it are identified. A patent may a) expand the existing patent lane in which the connected patent is located if its maximum semantic similarity exceeds a certain threshold value, e. g. the sum of the arithmetic mean and a single standard deviation or b) open up a new patent lane if the threshold value is not exceeded.
(v) The final step involves the extraction of keywords to characterize all patents as well as all lanes. For this purpose the TF IDF, as used for instance by Chen et al. (2012), can be applied. This measure outlines keywords that are specific for a single patent or patent lane but not for the complete set of patents. Furthermore, information from meta-data may be added.

4.2. Design decisions regarding patent lanes

The process of the development of patent lanes does not seem to be difficult in principle, but there are certain design decisions to be considered which influence the outcome of the process significantly. The most important design decisions concern the selection of the basic set and the starting set, the measurement of similarities, the threshold value to be

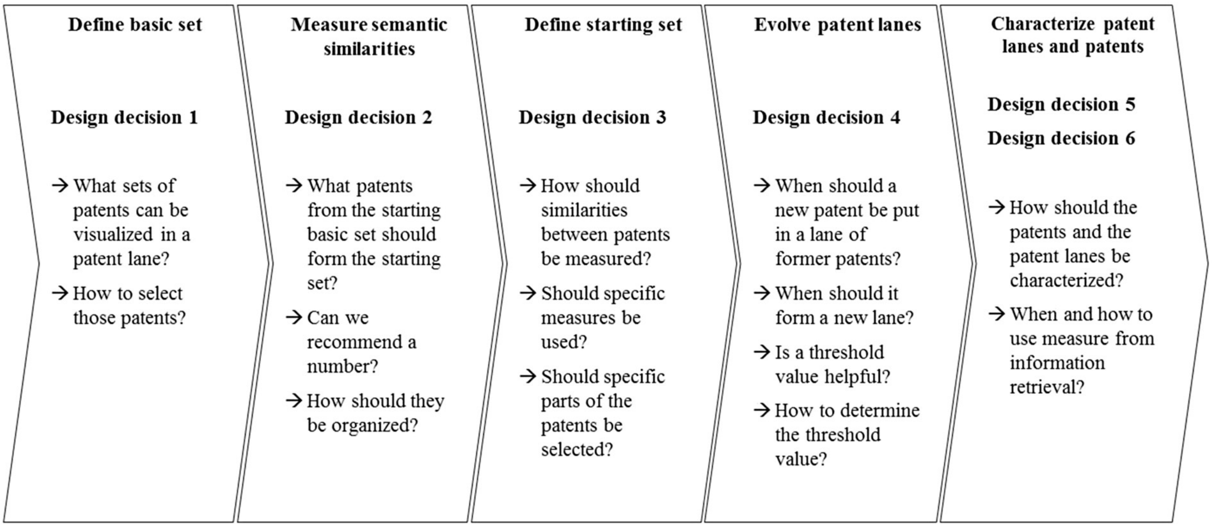

Define basic set
Design decision 1
→ What sets of patents can be visualized in a patent lane?
→ How to select those patents?
Measure semantic similarities
Design decision 2
→ What patents from the starting basic set should form the starting set?
→ Can we recommend a number?
→ How should they be organized?
Define starting set
Design decision 3
→ How should similarities between patents be measured?
→ Should specific measures be used?
→ Should specific parts of the patents be selected?
Evolve patent lanes
Design decision 4
→ When should a new patent be put in a lane of former patents?
→ When should it form a new lane?
→ Is a threshold value helpful?
→ How to determine the threshold value?
Characterize patent lanes and patents
Design decision 5
Design decision 6
→ How should the patents and the patent lanes be characterized?
→ When and how to use measure from information retrieval?

Fig. 2. Generic process for patent lane generation.
Source: authors.

used, the characterization of patents and patent lanes and the additional information provided by meta- data.

Design decision 1: Which sets of patents can be visualized in a patent lane? How are they selected (heuristics, auxiliary calculations)?

In analogy to classical cluster analysis the size of the patent set under research should be limited. We recommend a minimum of 15 patents and, for reasons of clarity, a maximum of 200 patents for a patent lane analysis. The patent set can be generated by using a search query for patents within a patent database, e. g. the United States Patent and Trademark Office (USPTO) full text application database (which covers patent applications) or the USPTO full text patent database (which covers granted patents). A query may for instance be specified by means of the patent classification, a keyword search, using patents by specific applicants, or combinations thereof (Alberts et al., 2011). If the initial search query yields more than 200 patents, there are two options that may help: The first option is to split the patent set into sub- sets by means of more deeply specified search queries, e. g. by focusing on IPC classes or CPC classes or keywords separately. The second option is to run a cluster analysis of the whole patent set and split it according to the emerging clusters.

Design decision 2: Which patents from the basic set should form the starting set? Is it possible to recommend a number? How should they be organized (should similarities between them be taken into account)?

There are two options for forming the starting set of patents. In the first and most simple one, the eldest patent alone forms the starting set of patents. We recommend this option if there are only a few patents in the basic set or if the patent activity within the technological field starts slowly or if both conditions apply. In the second option, the starting set of patents is formed by a sub-set of patents from the basic set. In contrast to the first option, we recommend this if there are many patents in the basic set or if the patent activity starts slowly or if both conditions apply. As a rule of thumb we suggest using up to ten patents to form the starting set. Some heuristics may be used to select the patents for the starting set, e. g. if there is any point in time at which the number of patents increases significantly or if other information of interest is given, such as the first co-classification of a patent in two different classes.

If the researcher chooses the second option involving more than one patent in the starting set, one or more cluster analyses may be run to find the starting lanes (see Kaufman and Rousseeuw, 2005 for methodical aspects of cluster analyses). Based on the similarity matrix, the patents of the starting set are subjected to a single linkage cluster analysis to identify possible outliers; those outliers should be examined separately, if they are really relevant. The remaining patents are subjected to a ward linkage cluster analysis. Finally, the relevant outliers and the clusters found by means of the ward linkage cluster analysis form the starting lanes.

Design decision 3: How should similarities between patents been measured? Should specific measures (Inclusion, Jaccard etc.) be used? Should specific parts of a patent be selected (e.g. claims, abstract)?

As explained in section two, there are different methods of semantic similarity measurement. Without going into detail, we recommend a method that has proven to be robust in our experience. In accordance with Niemann (2014) and Niemann and Moehrle (2013) we suggest using Flex N-grams with complete linkage and a Double-Single-Sided (DSS) inclusion calculation. In order to gain access to the information in the full text of a patent, four parts may be used for the similarity measurement, namely the description, the claims, the abstract and the title. Bibliographical and other meta-data should be excluded, because they only convey content related information indirectly, if at all. As a result of the semantic similarity measurement a similarity matrix can be obtained, displaying the similarities between all patents pairwise.

$$DI = \max\left(\frac{c_{i(j)}}{c_i};\frac{c_{j(i)}}{c_j}\right)$$

with the following variables:
DI = Double-Single-Sided Inclusion calculation
c_i = count of terms within document i
c_j = count of terms within document j
$c_{i(j)}$ = count of terms of document i that are also included in document j
$c_{j(i)}$ = count of terms of document j that are also included in document i.

Design decision 4: When should a new patent be included in a lane of a former patent, when should it form a new lane? Would a threshold value be helpful? How does one determine the threshold value (auxiliary calculation)? In which way is the threshold value related to the similarity measurement discussed above? In which way is the threshold value related to a prior structuring of the patents of a technological field e. g. by cluster analysis?

A new patent should be included in a lane of its antecedents if a certain semantic similarity exists, else it should constitute a new lane. A threshold value may be helpful in defining the necessary similarity (Niemann and Moehrle, 2013).[1] As the language of patents differs from technological field to technological field, and the values of the similarities vary in consequence of this, an absolute value for the threshold does not seem to be particularly helpful. Instead, statistical information may be used. We suggest starting with the sum of the arithmetic mean and a single standard deviation. If necessary, this threshold may later be changed. It is not only the language of the technological field under research which influences the similarity values, the chosen method of similarity measurement is decisive for those values as well. This is another reason for determining the threshold value in a case specific, appropriate way.

If the researcher has used a starting set with more than one patent and applied a cluster analysis, the results of this cluster analysis can be used to determine a threshold value in an alternative way. The distribution of similarity values in the clusters may be helpful, e. g. the minimum or the median might form a useful threshold value.

Design decision 5: How should patents and patent lanes be characterized? When and how can measures from information retrieval be used?

It seems to be useful to characterize both patents and patent lanes by means of keywords representing their distinctiveness and their novelty in the sense of exceeding prior art. For this purpose, researchers from the scientific field of information retrieval suggest a measure called "Term frequency – inverse document frequency" (Tf idf) (Chen et al., 2012). This measure outlines keywords which are specific for a single patent or patent lane, but not for the entire set of patents. The Tf idf consists of two components. The first component is the frequency of a specific term within a single document, in our case either a patent or a patent lane (seen as one document). The second component is the inverse document frequency. It is calculated by the logarithm of the quotient of the number of documents in a defined set divided by the number of documents containing the specific term.

$$Tf\ \ idf_{k,s} = (tf_{k,s}\ \cdot\ \log(\tfrac{c}{cf_k}))\forall_{k,s}$$

with the following variables:
tf = term frequency
idf = inverse document frequency
tf idf = term frequency inverse document frequency
k = index for terms
s = index for clusters
cf. = number of clusters containing the term.
c = sum of clusters.

Depending on the document under research (patent or patent lane) and on the defined set of documents forming the corpus, different options of extracting keywords based on the Tf idf present themselves[2]: (i) Patent specific keywords related to the patent lane: a specific patent lane is considered; each of its patents is compared against the corpus of former patents. This Tf idf demonstrates the chronological change of the technological perspective within a specific patent lane and can be interpreted as a proxy for novelty in the patent lane. (ii) Patent specific keywords related to the technological field: the entirety of patents of a technological field is considered. A certain patent of interest is compared against the corpus of all antecedent patents. This Tf idf points to the specific technological perspective within the technological field. (iii) Patent lane specific keywords related to the technological field: the entirety of patents of a technological field is considered. All patents of a patent lane are integrated in one single document which is compared against the corpus of all patents in the technological field. This Tf idf represents the specific topic of the patent lane at hand.

That a refinement may help to obtain more valid keywords[3] especially applies where patent lanes comprising a large number of patents are concerned. Once patent lane specific keywords have been calculated, they can be sorted according to their Tf idf values. As the keyword with the highest Tf idf does not necessarily cover all patents in the patent lane, additional keywords should be assigned. We suggest an iterative procedure. Starting with the keyword which has the highest Tf idf, all patents of a patent lane that contain this keyword are identified and deleted. After this, the keyword with the next highest Tf idf is selected. Again, all patents containing this keyword are identified. If a noteable number of patents show up, the keyword is added to the list of keywords which characterize this patent lane. In contrast, if the keyword is not contained in any of the remaining patents, the keyword is deleted. Subsequently, the patents containing this keyword are also deleted. The procedure continues with the keyword marked by the next highest Tf idf, until a specified number of keywords have been tested (we suggest using ten keywords) or until no patent of the set is left.

As an alternative or supplement to the use of Tf idf, the researcher could extract keywords manually, e. g. from the titles of the patents or the classification titles. This procedure requires a high manual effort (especially if used solely); on the other hand it may yield more valid results.

Design decision 6: What additional information may help to gain access to the technological field at hand? What meta-data (applicant, family status etc.) can be used?

It may be helpful to use additional information to enrich the information contained in the patent lanes and also to reorganize patent lanes. For instance, information about the applicants can be represented in different colors or some other graphical form to enable insights into the competitive structure of a technological field. Information about patent families or IPC or CPC classes may be used in the same way. The IPC or CPC classes or the applicants may also be employed to (re)organize the patent lanes. Classes or applicants may constitute major lanes, and the patent lanes related to the classes or the applicants are then assigned to these major lanes.

5. Case study in the field of carbon fiber reinforcements

Having introduced the basic process for patent lanes and discussed several design decisions, we now demonstrate the application of patent lanes by means of a case (see Yin, 2013 for selection criteria for case studies). We focus on an important topic in lightweight construction, the application of carbon fibers in bicycle technology (see Yang et al., 2012; Sargianis and Suhr, 2012 for an overview about application fields for carbon fibers). We select this case for the following five reasons: (i) We possess technical experience in this field, and are thus able to validate the results of our method ourselves. (ii) The case is easily assessable. (iii) The technology of carbon fibers has made significant progress over the past 20 years, but is still developing. (iv) The application field is driven by several new trends, and even though bicycle technology is well-established, newly emerging requirements have to be met. (v) As the US market is important for high-end bicycle producers, various US patents are available (which can be processed easily with our software tools). The case is characterized by a variety of companies which are active in patenting. Hence, we do not only expect to gain insight into the technology's development, but also into players' competitive strategies.

Our case is organized in six steps. First, we create a data set using US patents. Second, we measure semantic similarities between the patents

[1] Niemann (2014) used another heuristic. She arranged a patent in a new lane if the patent had at least two equal maximum similarities to its successors. She found that in such cases the similarities were rather low and used this as an additional argument for her heuristic.

[2] We concentrate on three ways of using Tf idf which we find most useful for patent lane analysis. Still other ways to apply Tf idf are possible.

[3] We thank the anonymous reviewer 1 for his suggestion to validate our dataset manually (see Appendix 1), which brought us to this refinement.

of our data set. Third, we define a starting set of patents. Fourth, we develop the patent lanes, and fifth, we characterize those lanes with an infometric approach. Sixth, we present and interpret our resulting patent lane diagram.

5.1. Creating the data set

As mentioned before, the United States Patent and Trademark Office (USPTO) provides two databases:

i. The USPTO Patent Full-Text and Image Database: This database contains granted full text patents dating from 1976 to the present, and images of granted patents from 1790 to the present.
ii. The USPTO Application Full-Text and Image Database: This database contains filed or applied patents dating from 2001 to the present.

The first decision is to use the Patent Full-Text database because the documents in this database contain full classification information which is provided by the patent examiner during the assessment process.[4] The second decision deals with the design of the search query: variable matching keywords are used to search for carbon fiber materials, e. g. fiber or fiber, carbon, composite, and reinforce. These keywords are searched in the abstract and title of the full text patent. To include different word combinations, truncations are used. Due to the link between carbon fiber materials and bicycles, the international patent classification (field code in the USPTO database: ICL) is used. The search query focusses on the technological field under research via bicycle related IPC subclasses: B60B, B62H, B62J, B62K, B62L and B62M. Furthermore, patents dating to the period from 1976 to 2014 are searched (19760101-20141130).

The final search query has the following form:

((ICL/(B60B$ OR B62H$ OR B62J$ OR B62K$ OR B62L$ OR B62M$) AND (TTL/(carbon AND TTL/(((fiber$ OR fibre$) OR composite$) OR reinforc$)) OR ABST/(carbon AND ABST/(((fiber$ OR fibre$) OR composite$) OR reinforc$)))) AND APD/1/1/1976->11/30/2014)

54 granted patents are found by using this search query.

5.2. Similarity measurement within the created data set

For a pairwise measurement of the semantic similarity among identified patents, the PatVisor®, a software tool developed at the Institute of Project Management and Innovation, is used. To compare the contents of the patents, bi-gram[5] concepts are extracted from a window of five concepts. In order to guarantee the validity of the similarity measurement specific filters are used during the analysis, e. g. stop word filters excluding common words like “and”, “or”, “the” or “a”; lemmatizers which transform verbs into their infinitive form; and patent related filters, for cleaning patent specific vocabularies. For calculating similarities between patents, a complete linkage setting is used in combination with a Double Single-Sided (DSS) Inclusion calculation.

5.3. Identifying starting set and deriving starting lanes from the data set

For the purpose of identifying the starting set, the ten oldest patents are chosen. The patents of the starting set were granted between 1982 and 1992. For deriving the starting lanes we proceed in two steps. First, in order to identify outliers, we execute a cluster analysis applying the single linkage method to the whole data set. Ten outliers can be found, four of them in the starting set. Second, we apply a ward cluster analysis to the remaining six patents of the starting set. This leads to a two cluster solution, in which cluster one contains four patents, and cluster two contains two patents. In total, six possible starting lanes emerge.

5.4. Evolving patent lanes from the data set

In the similarity matrix the 54 granted patents are arranged chronologically, in ascending order. During the next step the semantic similarity matrix is transformed into an upper triangular matrix, since this design is helpful for further analysis. The six starting lanes containing the ten patents of the starting set are marked. With the help of the threshold value the remaining 44 patents can be assigned to the correct starting lane with the right antecedent patent. To define the threshold value we used the sum of the arithmetic mean and a single standard deviation. The arithmetic mean of the data within the semantic similarity matrix is $\mu = 0,0621$, the standard deviation has the value $\sigma = 0,07821$. Therefore we calculate the threshold value as: $\mu + \sigma = 0,14031$. After calculating the threshold value we examine whether the successive patents exceed, or at least reach, the defined value. If the threshold value to an antecedent patent is reached or exceeded by a successive patent, the existing lane can be expanded. In eleven cases a successive patent exceeds the threshold value to only one single antecedent patent. If a successive patent exceeds the threshold value to several antecedent patents, only the maximum value is regarded. This happens to be the case twenty-three times; the respective patents have between two and thirteen antecedent patents. Ten patents do not reach the threshold value and thus open up new lanes. Finally, the 54 patents are arranged in 16 different patent lanes (see Fig. 3).

5.5. Retrieving information from the patents and patent lanes by use of keywords

After arranging the patents in different patent lanes, we extract keywords to characterize the patent lanes or the development within a patent lane over time. We calculate each patent lane's Tf idf to retrieve information about the patent lane's specific topics. Furthermore, we pick out patent lane fifteen to illustrate the use of the Tf idf within a specific patent lane.

To determine the different Tf idfs for the resulting sixteen patent lanes, four steps are necessary. (i) All patents of each patent lane are summarized into one larger document. Hence, the patent lane does no longer consist of different patents but of a single document only. (ii) All substantial bi-grams within the patent lanes document are extracted using the Term document matrix (TDM),[6] another analytic tool of the PatVisor®. (iii) The absolute number of a bi-gram represents the term frequency (Tf). Additionally, the inverse document frequency (idf) has to be calculated. We use the formula mentioned in design decision 5. (iv) After calculating the idf and joining it with the Tf, the resulting values are ranked in decending order. The bi-grams with the highest values are chosen to represent the patent lane.

In our case each of the sixteen patent lanes involves one or more unique term(s) which characterizes the topic of the patent lane (see Table 2). The use of carbon fibers within the technological field of the bicycle expands broadly. Several parts are reflected by the extracted keywords, e. g. seat post, pedal, frame or lock. Furthermore, some keywords are related to the material, e. g. metallic layer or vinyl/plasma compound. Even characteristics can be found: elastomeric/absorbing, fasten or tubular.

As mentioned before, a refinement concerning the extraction of keywords by use of the Tf idf may lead to more valid keywords describing patent lanes, especially large ones. The iterative refinement process is

[4] In contrast, patent applications need to carry only one classification which is suggested by the applicant himself or the patent attorney.

[5] Combinations of two words located both in a window of five concepts

[6] The PatVisor® sorts the concepts alphabetically; e. g. seat post turns into post seat. This should be kept in mind when viewing Tables 2 and 3.

Please cite this article as: Niemann, H., et al., Use of a new patent text-mining and visualization method for identifying patenting patterns over time: Concept, method and test ap..., Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.10.004

subsequently demonstrated in a case example. We select the largest patent lane – patent lane 5 – as a model. In the first round, this patent lane is named “post seat”, considering the Tf idf with the highest value only. Subsequently, we remove all patents from the patent lane which contain the term “post seat”. Next, the second highest Tf idf of this patent lane is considered: “steel tube”. There is only one patent containing this combination of words. We add this term to the lane description and delete the patent. In the next step, we take a look at the third highest Tf idf: “seat tube”. This term is comprised in three of the remaining patents; we add the term to the lane description and delete these patents as well. We continue to do so, until no further patents remain in patent lane 5. By means of this approach we generate a list containing a number of terms which properly describe the technological content of the patent lane (see Appendix 1). The new, refined keywords are “post seat/steel tube/seat tube/member tube/body unit/steer tube/assembly frame”. These combined Tf idfs actually describe patent lane 5 in its entirety much better than the single term “post seat” does.

To determine the change of the Tf idf within a certain patent lane over time, a method similar to the aforementioned one is used. The only difference between the keyword extraction from several patent lanes and an inner patent lane keyword extraction is that the corpus of patents expands: the title of the patent can be used to determine the topic at the patent lane's starting point. For identifying the Tf idf of the subsequent patent, a corpus has to be built. In this case, the corpus consists of the first patent and the successive patent. With the help of this corpus, the keyword or bi-gram can easily be extracted. If another patent is added to the patent lane at a later point, this new patent is attached to the corpus and expands it. The new Tf idf is calculated by using the new corpus and so on. This procedure can be repeated until a) the point of interest is reached or b) there are no new patents to be added to the patent lane.

For example, we focus on patent lane 15 “body tubular/fabric ply/fabric plurality/peripheral wall”. Four of this lane's patents were applied for on the same date (February 13th, 2002). In the three following years only one patent was applied for annually. Patent lane 15 contains seven patents in total. As the correct chronological order of the first four patents cannot be determined clearly, we define these four patents as a group-document. The Tf idf for each of the four patents is assigned using the four patents as the corpus and comparing the terms of each patent with the terms of the corpus one after the other. Furthermore, the corpus expands in the course of time. Whenever a new patent is applied for, it adds to the corpus. By means of this procedure, the Tf idf for each of the seven patents is defined (see Table 3).

5.6. Results

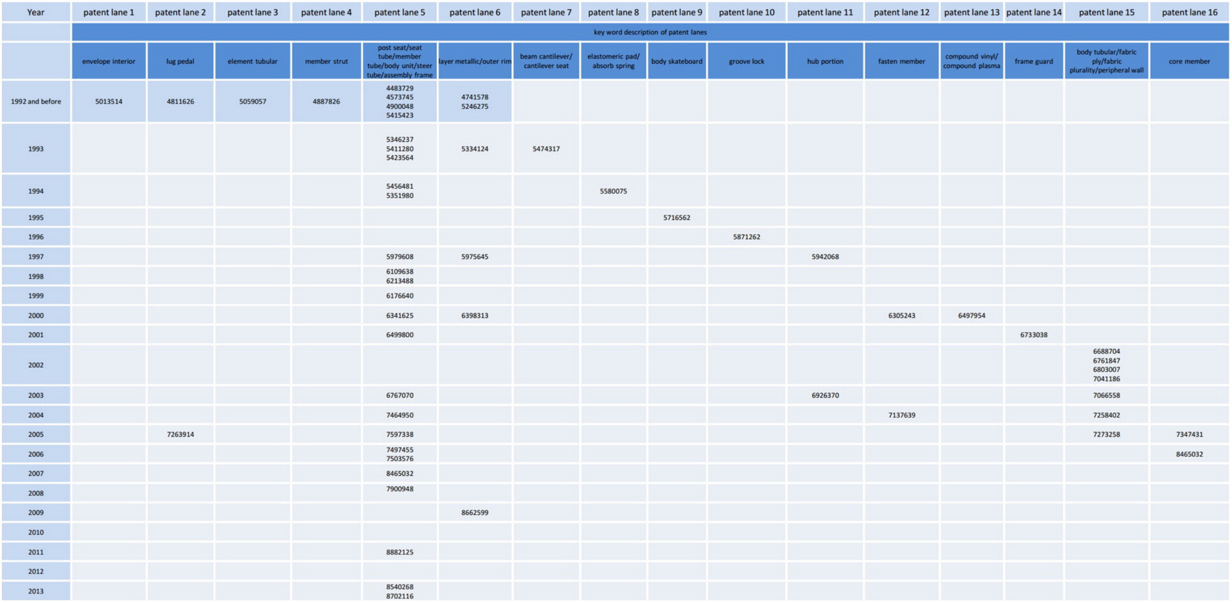
The resulting patent lane diagram is shown in Fig. 3 with raw data from PatVisor® including the patent numbers and in Fig. 4 in refined version. In total, sixteen patent lanes can be identified, covering different topics from the field of bicycle technology, e. g. seat post, groove lock or tubular body. Up to the year 1992 ten patents were applied for; these form our starting set. Further analysis leads to four outliers in this starting set. For three outliers no succeeding patents can be found, hence they are not represented in the diagram.

In the course of time three major lanes emerge: one dealing with the seat post, another with the metallic layer, and the third with the tubular body. While in the seat post lane and in the metallic layer lane the application activity was more or less constant over time, the application activity in the tubular body lane was concentrated in a short time frame between the years 2002 and 2005. As an interpretation of this we suggest that in the seat post lane and the metallic layer lane incremental invention is dominant, while in the tubular body lane a new (radical) topic has evolved and has been protected broadly, which is also indicated by the three different IPC classes to which the patents are assigned.

New lanes occur between 1983 and 2004, but no later than this. This indicates that carbon fibers had a major influence on bicycle technology at that time. Interestingly, the new lanes mainly consist of the one patent that constitutes them. Obviously, the carbon fibers affect different aspects of a bicycle without having a major impact on the technology as a whole.

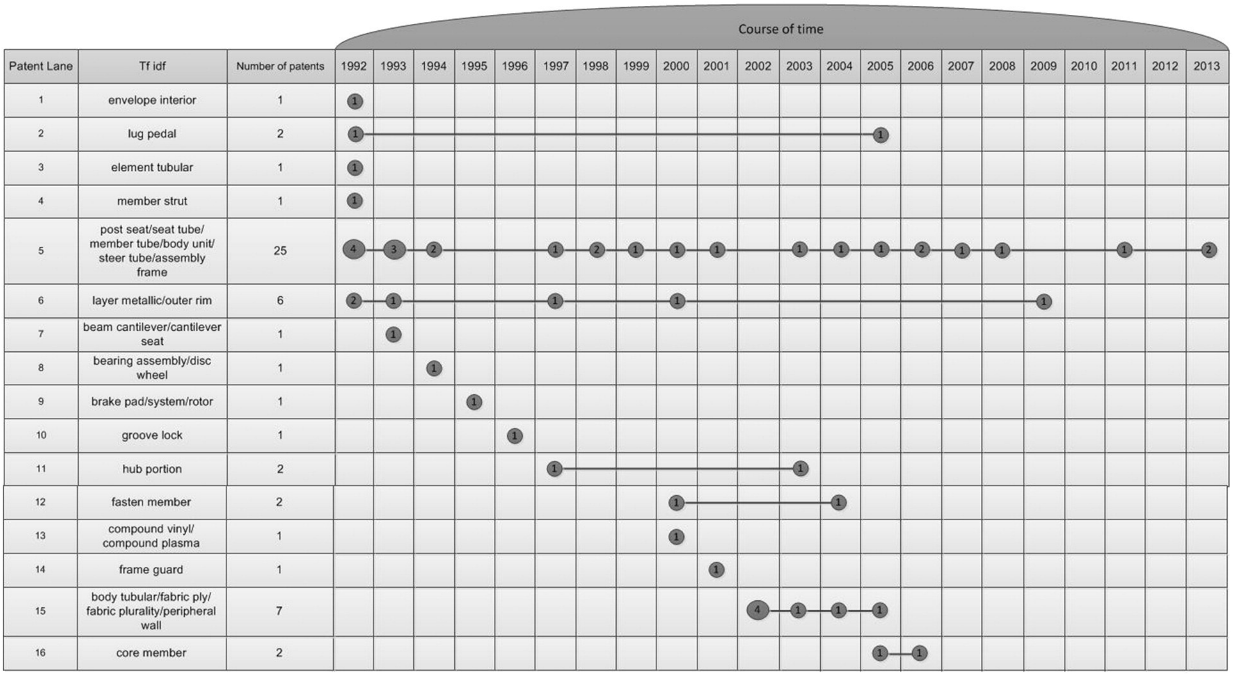
In addition to the basic patent lane diagram, there is Fig. 5 to be considered. It shows the links between patents, patent lanes and involved companies. Two types of observation become obvious: (i) company-activity-based ones, and (ii) patent lane-oriented ones. These enable even deeper insights into the technological field, its structure and its development.

(i) 24 companies are involved in the 16 patent lanes of the given technological field. Additionally, 13 inventors applied for their patents



Year	patent lane 1	patent lane 2	patent lane 3	patent lane 4	patent lane 5	patent lane 6	patent lane 7	patent lane 8	patent lane 9	patent lane 10	patent lane 11	patent lane 12	patent lane 13	patent lane 14	patent lane 15	patent lane 16
	key word description of patent lanes															
	envelope interior	lug pedal	element tubular	member strut	post seat/seat tube/member tube/body unit/steer tube/assembly frame	layer metallic/outer rim	beam cantilever/ cantilever seat	elastomeric pad/ absorb spring	body skateboard	groove lock	hub portion	fasten member	compound vinyl/ compound plasma	frame guard	body tubular/fabric ply/fabric plurality/peripheral wall	core member
1992 and before	5013514	4811626	5059057	4887826	4483729 4573745 4900048 5415423	4741578 5246275										
1993					5346237 5411280 5423564	5334124	5474317									
1994					5456481 5351980			5580075								
1995									5716562							
1996										5871262						
1997					5979608	5975645					5942068					
1998					6109638 6213488											
1999					6176640											
2000					6341625	6398313						6305243	6497954			
2001					6499800									6733038		
2002															6688704 6761847 6803007 7041186	
2003					6767070						6926370				7066558	
2004					7464950							7137639			7258402	
2005		7263914			7597338										7273258	7347431
2006					7497455 7503576											8465032
2007					8465032											
2008					7900948											
2009						8662599										
2010																
2011					8882125											
2012																
2013					8540268 8702116											

Fig. 3. Patent lane diagram for carbon fibers applied in bicycle technology, raw data from PatVisor® including patent numbers.
Source: authors.



			Course of time																					
Patent Lane	Tf idf	Number of patents	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
1	envelope interior	1	1																					
2	lug pedal	2	1													1								
3	element tubular	1	1																					
4	member strut	1	1																					
5	post seat/seat tube/ member tube/body unit/ steer tube/assembly frame	25	4	3	2			1	2	1	1	1		1	1	1	2	1	1			1		2
6	layer metallic/outer rim	6	2	1				1			1									1				
7	beam cantilever/cantilever seat	1		1																				
8	bearing assembly/disc wheel	1			1																			
9	brake pad/system/rotor	1				1																		
10	groove lock	1					1																	
11	hub portion	2						1						1										
12	fasten member	2									1				1									
13	compound vinyl/ compound plasma	1									1													
14	frame guard	1										1												
15	body tubular/fabric ply/ fabric plurality/peripheral wall	7											4	1	1	1								
16	core member	2														1	1							

Fig. 4. Patent lane diagram for carbon fibers applied in bicycle technology, refined visualization.
Source: authors.

individually. As the individual inventors cannot be associated with any specific company, they do not qualify as research objects. Four statements can be made regarding the 24 companies:

- There are 14 companies which only hold one single patent.
- Seven companies possess two patents each. Six of these companies hold their patents in a single patent lane; their patent activities are mono-fractioned. One company holds one patent in patent lane 2 and another patent in patent lane 6.
- Two companies own three patents each. Their patent activities are mono-fractioned.
- One of the companies holds eight patents. Seven of these patents form patent lane 15. This patent lane completely belongs to the company. Another one of this company's patents can be found in patent lane 12.

(ii) The patent lanes are either mono-fractioned or multi-fractioned. For instance, patent lanes 15 and 16 are mono-fractioned. The patents included in these patent lanes have a single origin, i.e. all seven patents of patent lane 15 are Campagnolo-patents. Patent lane 5 describes the opposite pole; it is multi-fractioned. The 25 patents in this patent lane were filed by twelve different companies.

The analyses provide an impression of the companies' strategic patenting activities. Basically, there are three different types of strategic patenting:

- Selective-dominant patenting: Campagnolo can be seen as a prime example of this strategy. This company dominantly claims an entire patent lane, thus defining the state of the art in this field. Campagnolo filed their patents between 2002 and 2005; obviously the company had a breakthrough in a specific part of bicycle technology in this period.
- Selective-emphasized patenting: Other companies emphasize their patent activities in existing patent lanes without being dominant. For instance, Softride set an emphasis in patent lane 5. This company is an early player in the field of seat posts and filed its patents between 1992 and 1994. It is similar with Trek Bicycles, which hold a follower position; they filed their patents between 1998 and 2008.
- Sporadic patenting: The remaining companies file patents sporadically. By way of single inventions they fill in occasional white spots in existing patent lanes.

6. Application profile of patent lanes compared with rolling clustering

Having provided a proof of concept with our case study, we now aim at profiling our method. For this purpose we select an alternative method class for analyzing evolutionary patterns of technological fields based on rolling clustering. We choose this method class because, in accordance with our method, it is targeted at disaggregating a technological field over time. Below, five methods for the use of rolling clustering are listed.

Upham et al. (2010a) analyze innovating knowledge communities with the help of conference and journal papers. They use a new clustering approach called StrEMer which produces high-quality clusters that are dynamic[7] in the course of time. They organize their data in overlapping five year periods and receive clusters by years. This approach enables a backward-looking view on all cluster assignments from each year within a five year time frame.

Upham et al. (2010b) analyze how new knowledge is created. For this purpose they use articles from top journals issued between 1956 and 2002 as their dataset. A co-citation analysis is used to establish the network structures between the research papers. Again, Upham et al. (2010b) use StrEMer for generating clusters in a rolling way. The time frame is split into 10-year blocks starting at annual intervals, e. g. the clustering of the 1990 data is based on elements from 1980 to 1990, the 1991 data is based on elements from 1981 to 1991 and so forth (see Upham et al., 2010b).

Denny et al. (2010) visualize temporal cluster changes using relative density self-organizing maps. They use different world development

[7] Other current computer-implemented clustering algorithms consider clusters as static in the course of time (see Upham et al., 2010b). This might not lead to errors regarding short time slices, but is not productive regarding large time slices.

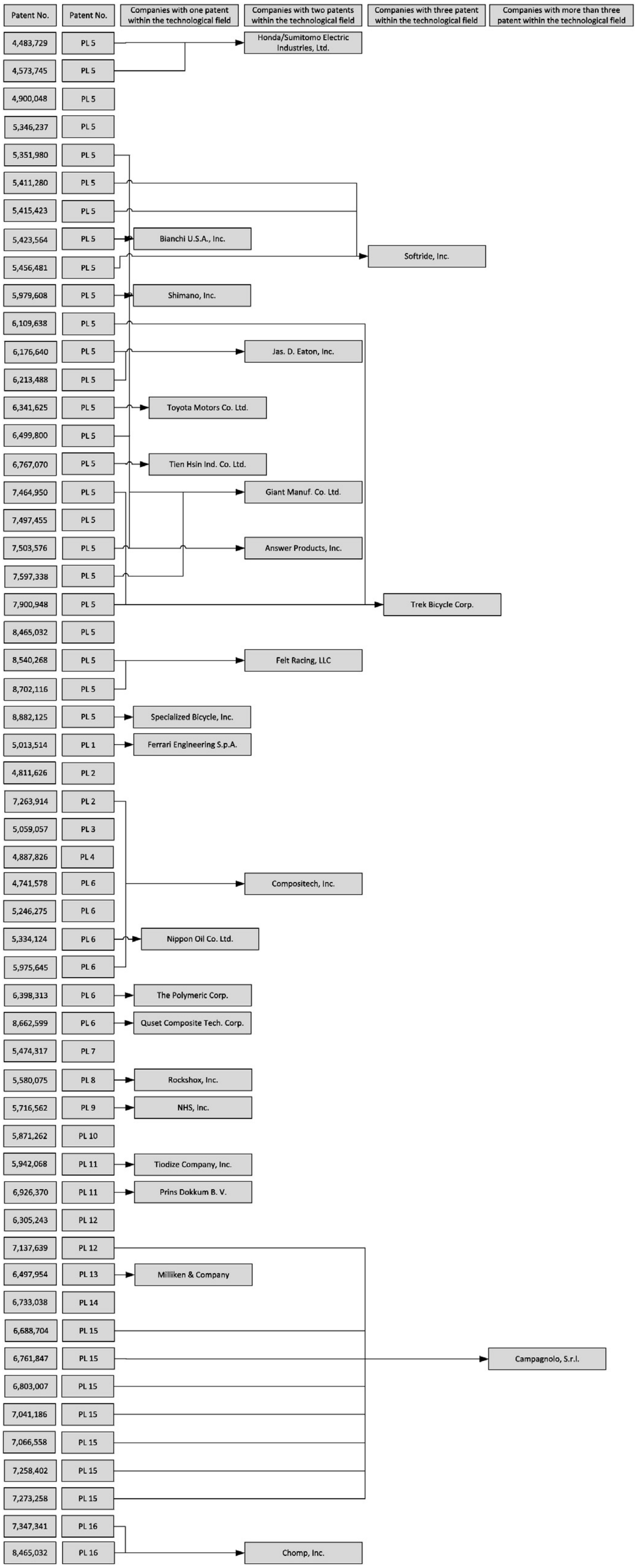


Fig. 5. Patent lanes combined with assignees for carbon fibers applied in bicycle technology. Source: authors.

Table 2
Results of the keyword extraction for all 16 patent lanes, using the Tf idf and the refinement procedure. Source: authors. Remark: The PatVisor® lists the terms in alphabetic order, so for instance "seat post" is extracted as "post seat".

Patent lane	Keywords
1	"envelope interior"
2	"lug pedal"
3	"element tubular"
4	"member strut"
5	"seat post"
6	"metallic layer"
7	"cantilever beam" or "cantilever seat"
8	"elastomeric pad" or "absorb[*ing*] spring"
9	"body skateboard"
10	"groove lock"
11	"hub portion"
12	"fasten member"
13	"vinyl/plasma compound"
14	"frame guard"
15	"tubular body"
16	"core member"

indicators as their data set. Applying the relative density self-organizing maps in combination with distance matrices and color linking, different changes within clusters can be visualized: (i) emerging clusters, (ii) disappearing clusters, (iii) split clusters, (iv) merged clusters, (v) enlarging clusters, (vi) contracting clusters, (vii) the shifting of cluster centroids as well as (viii) changes in cluster density. For analyzing the clusters different time slices were used.

Shibata et al. (2010) identify a commercialization gap between science and technology in their paper. For this purpose they use a citation network analysis which is based on (a) patent data to define the technological side and (b) publications to describe the scientific side. Within the citation network analysis a topological clustering method was used. The results were visualized along a timeline. By comparing the development of the two sides (science and technology) white spots can be marked and exploited commercially. In this method clusters are arranged over time; they are not further disaggregated to the patent level.

Small (2006) describes the growth areas in science by means of a co-citation cluster analysis. He uses frequently cited papers for data input. To determine the growth of the research area, Small (2006) chooses a dataset encompassing a period of six years. He divides this six-year frame into three time slices to clarify the emergence of the field in a more detailed way.

All of the above-mentioned methods have two aspects in common: they disaggregate a technological field into clusters, and they mostly involve overlapping time slices for clustering. Although these methods are useful in general, some differences between them and our method point to differing application profiles: (i) Our method seems to have an advantage if the starting point of a technological field can be identified as lying in the near past (e.g. if it is an emerging technology). Otherwise, especially if the starting point is located in the more distant past, and if a lot of patents can be assigned to the given technological field, the patent lane diagram may spread widely and thus become very broad and

Table 3
Results of the keyword extraction for all patents in patent lane 15 "tubular body". Source: authors.

Patent	Application date	Keywords
6,688,704	02/13/2002	"central axis"
6,761,847	02/13/2002	"peripheral wall"
6,803,007	02/13/2002	"main portion" or "bicycle frame"
7,041,186	02/13/2002	"tubular body"
7,066,558	07/08/2003	"longitudinal axis"
7,258,402	04/01/2004	"flange wall" or "wall flange"
7,273,258	06/10/2005	"axis fiber" or "arrangement overlap" or "hollow proximate"

confusing. In such cases rolling clustering methods can be more useful. (ii) Our method has an advantage if the analysts aim at continuously monitoring a technological field rather than at performing a single analysis. If a new patent application is disclosed, it can easily get integrated into the patent lanes without the necessity of changes in previous structures or previously assigned Tf idfs. (iii) Using Tf idfs in different variants in our method allows us to identify novel aspects in the patents and patent lanes of the technological field under research. At the moment this seems to be a unique feature.

7. Conclusions

In this paper we have introduced patent lanes, which we define as the deployment of patent clusters in the course of time. For this purpose we have developed a generic process consisting of five steps, and discussed six design decisions related to these steps. The field of carbon fibers in combination with bicycle technology was chosen to demonstrate the patent lane method, leading to insights into the evolution of the technological field over time.

Our method has several limitations. As always, when using semantic similarity as a proxy for content similarity, the question arises how good both concepts are in relation to each other. A bias may be caused by various patent attorneys using different terms for the same concept, by different explanations of an invention, etc. Second, the technological field under research may play a role, as in some fields, such as chemistry and pharmaceutics, semantic analyses are complicated by the fact that a chemical formula can be written in various ways. Third, the strategic thinking of inventors and patent attorneys may have an influence on our analyses (in addition to accidental peculiarities of language use); as some of them may deliberately employ diction that is difficult to analyze. Forth, the way in which we define the threshold value represents a further limitation of our method. It may be helpful to use other statistical measurements, e. g. the median or technological field specific maximum or minimum values. Fifth, as regards obtaining a starting set we have already discussed two different approaches, i.e. cluster analysis or an iterative method starting with the technological field's first patent. There may be other approaches that might also influence the results of our method. Sixth, we use the Tf idf measure to retrieve keywords for patent lanes and patents. As Zhang et al. (2011) point out; there may be more appropriate measures to achieve this.

In connection with some of the stated limitations, we have already suggested a few questions for further research. In addition to these, there is one methodical aspect that seems to be worth further investigation: the assignment of a successive patent to an antecessor patent. For reasons of manageability and simplicity, we use a solitary assignment in our method: A successive patent is assigned to only one of the antecessor patents, even if the threshold to more than one is reached or even exceeded. Relaxing this rule would lead to a network instead of a tree diagram. By additionally reducing the threshold value, a more and more connected, intermeshed network emerges. This kind of network would enable analyses similar to those of the patent citation networks.

All in all, patent lanes may become an important method for scientometric purposes. They provide deep insights into the evolution of a technological field by using measurable attributes. Those insights could be useful for analysts in research institutions as well as for practitioners in companies. The underlying semantic similarity measurement provides an alternative to classical citations. It offers some advantages regarding timely availability, for instance, but has to be applied with great care in order to avoid or reduce a possible bias, as discussed in the limitations paragraph. As a further perspective, semantic similarity measures could also be used in combination with citations (as already used by Liu et al., 2010), resulting in a hybrid approach with the capability to improve network analyses.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.techfore.2016.10.004.

References

Alberts, D., Yang, C.B., Fobare-DePonio, D., Koubek, K., Robins, S., Rodgers, M., 2011. Introduction to patent searching. In: Lupu, M., Mayer, K., Tait, J., Trippe, A.J. (Eds.), Current Challenges in Patent Information Retrieval. Springer, Heidelberg, pp. 3–44.

Boyack, K.W., Klavans, R., 2010. Co-citation analysis, bibliographic coupling, and direct citation: which citation approach represents the research front most accurately? J. Am. Soc. Inf. Sci. Technol. 61 (12), 2389–2404.

Chang, S., Lai, K., Chang, S., 2009. Exploring technology diffusion and classification of business methods: using the patent citation network. Technol. Forecast. Soc. Chang. 76, 107–117.

Chen, S., Huang, M., Chen, D., 2012. Identifying and visualizing technology evolution: a case study of smart grid technology. Technol. Forecast. Soc. Chang. 79, 1099–1110.

Choi, C., Kim, S., Park, Y., 2007. A patent-based cross impact analysis for quantitative estimation of technological impact: the case of information and communication. Technol. Forecast. Soc. Chang. 74, 1296–1314.

de Souza Carvalho, D., Guimarães de Oliveira, L., Winter, E., 2009. Technological foresight based on citing and cited patents of cellulose with pharmaceutical applications. J. Technol. Manag. Innov. 4, 32–41.

Denny, Graham, J.W., Christen, P., 2010. Visualizing temporal cluster changes using relative density self-organizing maps. Knowl. Inf. Syst. 25, 281–302.

Dereli, T., Durmusoglu, A., 2009. A trend-based patent alert system for technology watch. J. Sci. Ind. Res. 68 (8), 674–679.

Ernst, H., 2003. Patent information for strategic technology management. World Patent Inf. 25, 233–242.

Frietsch, R., 2007. Patente in Europa und der Triade - Strukturen und deren Veränderung. Fraunhofer Institut für System- und Innovationsforschung, Karlsruhe.

Gerken, J.M., Moehrle, M.G., 2012. A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis. Scientometrics 91 (3), 645–670.

Gilsing, V., Nooteboom, B., Vanhaverbeke, W., Duysters, G., van den Oord, A., 2008. Network embeddedness and the exploration of novel technologies: technological distance, betweenness centrality and density. Res. Policy 37, 1717–1731.

Glänzel, W., Thijs, B., 2012. Using 'core documents' for detecting and labelling new emerging topics. Scientometrics 91 (2), 399–416.

Gower, J.C., Legendre, P., 1986. Metric and Euclidean properties of dissimilarity coefficients. J. Classif. 3 (1), 5–48.

Jaffe, A.B., Trajtenberg, M., Henderson, R., 1993. Geographic localization of knowledge spillovers as evidence by patent citations. Q. J. Econ. 108, 577–598.

Jeong, B., Lee, D., Cho, H., Lee, J., 2008. A novel method for measuring semantic similarity for XML schema matching. Expert Syst. Appl. 34 (3), 1651–1658.

Karvonen, M., Kässi, T., 2013. Patent citations as a tool for analysing the early stages of convergence. Technol. Forecast. Soc. Chang. 80 (6), 1094–1107.

Kaufman, L., Rousseeuw, P.J., 2005. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, Hoboken, New Jersey.

Kayal, A., 1999. Measuring the pace of technological progress. Technol. Forecast. Soc. Chang. 60 (3), 237–245.

Kuusi, O., Meyer, M., 2007. Anticipating technological breakthroughs: using bibliographic coupling to explore the nanotubes paradigm. Scientometrics 70, 759–777.

Lai, K.K., Wu, S.J., 2005. Using the patent co-citation approach to establish a new patent classification system. Inf. Process. Manag. 41 (2), 313–330.

Lee, S., Yoon, B., Lee, C., 2009. Business planning based on technological capabilities: patent analysis for technology-driven roadmapping. Technol. Forecast. Soc. Chang. 76, 769–786.

Lee, H., Lee, S., Yoon, B., 2011. Technology clustering based on evolutionary patterns: the case of information and communications technologies. Technol. Forecast. Soc. Chang. 78, 953–967.

Lee, C., Cho, Y., Seol, H., 2012. A stochastic patent citation analysis approach to assessing future technological impacts. Technol. Forecast. Soc. Chang. 79, 16–29.

Liu, X., Yu, S., Janssens, F., Glänzel, W., Moreau, Y., De Moor, B., 2010. Weighted hybrid clustering by combining text mining and bibliometrics on large-scale journal database. J. Am. Soc. Inf. Sci. Technol. 61 (6), 1105–1119.

Moehrle, M.G., 2010. Measures for textual patent similarities: a guided way to select appropriate approaches. Scientometrics 85 (1), 95–109.

Moehrle, M.G., Gerken, J.M., 2012. Measuring textual patent similarity on the basis of combined concepts: design decisions and their consequences. Scientometrics 91 (3), 805–826.

Möller, A., Moehrle, M.G., 2015. Complementing keyword search with semantic search – introducing an iterative semiautomatic method for near patent search based on semantic similarities. Scientometrics 102 (1), 77–96.

Niemann, H., 2014. Corporate Foresight mittels Geschäftsprozesspatenten. Springer, Berlin et al.

Niemann, H., Moehrle, M.G., 2013. Car2X-Communication mirrored by business method patents: what documented inventions can tell us about the future. Technology Management in the IT-Driven Services. Proceedings of PICMET 2013.

Opsahl, T., Agneessens, F., Skorvetz, J., 2010. Node centrality in weighted networks: generalizing degree and shortest paths. Soc. Networks 32, 245–251.

Rost, K., 2011. The strength strong ties in the creation of innovation. Res. Policy 40, 588–604.

Sargianis, J., Suhr, J., 2012. Core material effect on wave number and vibrational damping characteristics in carbon fiber sandwich composites. Compos. Sci. Technol. 72, 1493–1499.

Shibata, N., Kajikawa, Y., Sakata, I., 2010. Extracting the commercialization gap between science and technology – case study of solar cell. Technol. Forecast. Soc. Chang. 77, 1147–1155.

Small, H., 2006. Tracking and prediction growth areas in science. Scientometrics 68 (3), 595–610.

Sorenson, O., Rivkin, J.W., Fleming, L., 1993. Compexity, networks and knowledge flow. Res. Policy 33, 994–1017.

Tseng, F., Hsieh, C., Peng, Y., 2011. Using patent data to analyze trends and the technological strategies of the amorphous silicon thin-film solar cell industry. Technol. Forecast. Soc. Chang. 78, 332–345.

Upham, S.P., Rosenkopf, L., Ungar, L.H., 2010a. Innovating knowledge communities. Scientometrics 83, 525–554.

Upham, S.P., Rosenkopf, L., Ungar, L.H., 2010b. Positioning knowledge: schools of thought and new knowledge creation. Scientometrics 83, 555–581.

von Wartburg, I., Teichert, T., Rost, K., 2005. Inventive progress measured by multi-stage patent citation analysis. Res. Policy 34, 1591–1607.

Yang, Y., Boom, R., Irion, B., van Heerden, D.-J., Kuiper, P., de Wit, H., 2012. Recycling of composite materials. Chem. Eng. Process. Process Intensif. 51, 53–68.

Yin, R.K., 2013. Case Study Research: Design and Methods. Sage, Thousand Oaks et al.

Yoon, J., Kim, K., 2011. Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks. Scientometrics 88, 213–228.

Zhang, W., Yoshida, T., Tang, X., 2011. A comparative study of TF*IDF, LSI and multi-words for text classification. Expert Syst. Appl. 38 (3), 2758–2765.

Helen Niemann was research associate at the IPMI - the Institute of Project Management and Innovation – between 2010 and 2014. Her research interests are focused on business method patents and corporate foresight. Between 2005 and 2010 she studied Business Administration at the University of Bremen.

Martin G. Moehrle has been the director of the Institute for Project Management and Innovation (IPMI) at the University of Bremen since 2001 while at the same time holding the chair of Innovation and Competence Transfer. His area of work is technology and innovation management, and his preferred topics are patent management, technology roadmapping, future research, innovation process management, and methodical invention (TRIZ). He obtained his doctorate and qualified as a university lecturer on transitional themes between technology management and business informatics.

Jonas Frischkorn is a PhD candidate in patent management at the University of Bremen. He holds an MSc in Engineering and Management from the University of Bremen. His work and research interests include particularly patent analysis, IP strategies in general, technology foresight, and competitive analysis.

Please cite this article as: Niemann, H., et al., Use of a new patent text-mining and visualization method for identifying patenting patterns over time: Concept, method and test ap..., Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.10.004