World Patent Information 50 (2017) 64–72

Contents lists available at ScienceDirect

World Patent Information

journal homepage: www.elsevier.com/locate/worpatin

Claim-based patent indicators: A novel approach to analyze patent content and monitor technological advances

Douglas Henrique Milanez [a, b, *], Leandro Innocentini Lopes de Faria [a, b], Roniberto Morato do Amaral [a, b], José Angelo Rodrigues Gregolin [a]

[a] Federal University of São Carlos, Centre for Technological Information in Materials, Dept. of Materials Engineering, 13565-905, São Carlos, SP, Brazil
[b] Federal University of São Carlos, Centre for Technological and Organizational Intelligence Research, Dept. of Science Information, 13565-905, São Carlos, SP, Brazil

ARTICLE INFO

Article history:
Received 20 December 2016
Received in revised form
8 August 2017
Accepted 21 August 2017
Available online 19 September 2017

Keywords:
Bibliometrics
Nanotechnology
Nanomaterials
USPTO
Innovation forecasting

ABSTRACT

This paper proposes a new method for developing patent indicators by text mining patent claims according to their drafting structure. We apply the method on nanocellulose as a case of study, although any subject could be the target of investigation. The results show that patent claims are a more reliable source of key terms to develop technical indicators than, for example, patent titles and abstracts. Indicators from patent claims in combination with other traditional indicators developed from bibliographic patent data may contribute significantly to the analytical process of technological forecasting, monitoring and competitive intelligence studies.

1. Introduction

1.1. Patents as a source of technological indicators

Patent statistics have been frequently adopted to follow technological trends, innovative activities, market analysis, and players involved in the innovation process, at least since the second half of the last century. In 1985, Pavitt [1] had already noticed the increasing use of patent statistics and he associated it to the growing recognition of the importance of technological change in firm competitiveness, improvements in technologies of information storage and retrieval, and the need for statistical evidence to support personal experience and expert opinion. The high volume of data currently available also justifies the need for quantitative analysis based on patent information [2].

Patent documents are rich sources of technological and business information. They codify part of the tacit knowledge generated by the technological development of new products, processes, methods, compositions, apparatus, etc. They also enable reasonably standardized quantitative data to perform accurate assessments using data and text mining approaches [2–4]. These assessments allow evaluating the effects of trade and industry production, development of industrial sectors, policy making related to scientific technological activities, and links between science and technology [1–3,5]. Much research and reports on using patent indicators can be found in the literature [for instance, see Refs. 3–13]. Moreover, Madani and Weber [4] performed an interesting review of patent mining evolution using bibliometrics and keyword network analysis.

In the context of technological forecasting, many studies have been conducted using patents as technological sources for quantitative evaluation. Considering patent analysis as a means of investigating phases of the technology life cycle, Chachetti et al. [14] overviewed the technological activity in hydrogen storage materials using patent indicators and network analysis for gathering insights on the future developments in this field. Caviggioli [11] investigated technology fusion by verifying convergence relying on the International Patent Classification (IPC) of patents filed at the European Patent Office between 1991 and 2007. His hypothesis was that the first occurrence of a patent incorporating a combination of IPC subclasses signals a new instance of fusion.

* Corresponding author. Federal University of São Carlos, Centre for Technological Information in Materials, Dept. of Materials Engineering, 13565-905, São Carlos, SP, Brazil.
E-mail address: douglasmilanez@yahoo.com.br (D.H. Milanez).

http://dx.doi.org/10.1016/j.wpi.2017.08.008
0172-2190/© 2017 Elsevier Ltd. All rights reserved.

Another example is the approach for forecasting promising technology proposed by Kim and Bae [12], where forward citations, triadic patent families and independent claims were used to assess promising technology clusters.

Most of the indicators are assembled by mining the bibliographic data of patent documents, which usually includes the IPC codes, priority numbers, countries of filing, inventors and patentees, etc. [15]. In more specialized studies, citations to prior patents and non-patent literature have also been used to improve the data mining exploitation of patent documents [16]. Van Raan [3] provided a detailed and insightful review of patent citation analysis and a new approach to map technology-relevant areas, focusing on patent-patent citation in the context of economic value of patents as well as citation to non-patent literature to landscape S&T linkage. One of the main conclusions is that only 3–4% of scientific publications covered by Web of Science or Scopus are cited by patents.

Non-controlled free text fields, such as titles and abstracts, have also been mined in order to extract key-terms that could depict the technical content of a set of patent documents. Courtial, Callon and Sigogneau [9] applied co-word analysis to normalized terms extracted from patent titles to get a panorama of a given field. The work of Brietzman [8] is another example of mining terms from patent titles and abstracts to assess industry R&D. Nevertheless, the use of titles and abstracts as a source for term extraction is limited and cannot provide all relevant aspects of an invention [17,18].

To fill this gap, researchers have been exploiting the content from the full text patent documents. For instance, key-terms have been extracted with routines based on the morphological and syntactic structure of sentences [17–26]. Other more specific techniques rely on segmenting and combining terms [17,21,26,27], summarizing before extracting key-words [17], vector representation and grouping approaches [28–30], clustering analysis [7,17,30] and mapping [22,25,26,29]. Nonetheless, less attention has been paid to considering rules for the drafting of each part of the patent document, such as the specific rules dictating the writing of patent claim sentences.

1.2. Patent claims as a source of indicators

Patent claims can be understood as “the heart of a patent” because it specifies the invention's scope of protection. According to the United State Patent and Trademark Office (USPTO) [31], claims must point out “the subject matter that the inventor or inventors regard as the invention, […] (and what) defines the scope of the protection of the patent”. Thereby, claims are a valuable source of technical terms [32]. To extract useful key-words from the claims, several methodologies have been proposed with different aims, such as improvement of wording and translation during the filing period of a patent document [19,33], information retrieving [34], claim overlapping and legal analysis [22,35], conceptual mapping [22], and technological forecasting and competitive intelligence [23–26]. Nevertheless, none of these studies dealt properly with the particularities of the patent claim structure. For instance, claim sentences can be relatively long, reaching more than 200 words, which can cause failures in the natural language processing [36].

According to the WIPO Patent Drafting, a claim can be divided into three main parts [32]: 1) the preamble, which introduces the category of invention or an essential part of the invention – can be a process, a method, a composition, a product, etc.; 2) the transition phrase, which separates the preamble from the body of the claim; and 3) the body of the claim, which details the characteristics of the invention. Patent claims are also divided into independent claims, which contain the general aspect of the invention, and dependent claims, which details each part of the invention – and may always be referred directly or indirectly to an independent claim, i.e. all dependent claims should be grouped together with the independent claim(s) to which they refer to [31,32].

The claim's “internal structure” can be exemplified with two sentences present in a patent document, as shown in Table 1. These two sentences are the first and the second claims from the patent document numbered US20110086948 [37] and we have already separated them according to the structure of writing. The first sentence is an independent claim while the second one is a dependent claim. Both preambles refer to a product (composite material) and the body of these claims contains useful key words that details the referred invention: “nanocellulose”, “maleic anhydride graft poly(ethylene-octene) copolymer resin”, “nylon-4 resin”, etc. In the approach that we propose, we considered the claim structure in the text mining routine that will be described in this paper and it allowed us to extract more accurately detailed information regarding products, compositions and processes protected by patents.

This paper presents a new method for developing patent indicators using patent claims as a source of key-terms. The idea is to investigate technical details in order to provide technological indicators with high aggregate value thereby supporting technological forecasting studies and decision making. The method take into account the structure of drafting the claim, and it was designed to analyze patents filed in the United States Patents and Trademarks Office (USPTO), however the logical principle of depicting the claim structure can be applied to patent documents in other offices. We used nanocellulose as a case study to delineate a sample of full-text patent documents, but it can be used for any other subject or technological field.

1.3. Nanocellulose: a sustainable nanomaterial

The technological developments of nanotechnology and nanomaterials have grown at least since the beginning of this century. High budgets to support research and development in these topics have been executed due to their great potential in promoting innovations. Consequently, nanotechnology has been the target of monitoring and forecasting activities based on patent indicators [for instance, see Refs. 38–44]. Nanocellulose is an emerging and economically promising nanomaterial that can be obtained from renewable sources, such as plants, woods, natural fibers, etc. It has been estimated that the American market for nanocellulose in 2020 will be US$ 250 million and that its production can reach 780 tons in 2017 [45].

The mechanical properties of nanocellulose are higher than the ones from conventional cellulose fibers. Furthermore, the attractive properties of nanocellulose also include biocompatibility and biodegradability, gas barrier, water absorption and rheological and optical properties. Among the main applications, besides being a reinforcing agent in composite materials and paper, we can mention packaging, optically transparent paper for electronic devices, texturizing agents in cosmetics and food, dressings and bioartificial implants [46–51].

Nanocellulose is a generic term for a set of cellulose-based nanomaterials, which also include cellulose nanofibrils, cellulose nanocrystals, bacterial cellulose [18,47,49,51,52]. In the top-down manufacturing process, nanostructures are obtained by

Table 1
Example of patent claims divided according to their text structure.

Preamble	Transition phrase	Body of the claim
1) A nylon-4 composite	comprising:	a nanocellulose; and a maleic anhydride graft poly(ethylene-octene) copolymer resin
2) The nylon-4 composite	of claim 1, wherein	the nanocellulose is dispersed in the maleic anhydride graft poly(ethylene-octene) copolymer resin, and the resulting maleic anhydride graft poly(ethylene-octene) copolymer resin is dispersed in the nylon-4 resin

diminishing the particle size of cellulose sources using mechanical approaches (nanofibrils), acid hydrolysis (nanocrystals) or a combination of them. In the bottom-up manufacturing process, nanostructures are formed by bacterial fermentation of carbohydrates and alcohols [47,48]. There are other synonymous with the prefix “micro” (such as microfibrillated cellulose), but they may be considered nano-sized materials as well, according to experts [49,52]. By following the patent activity from each type of nanocellulose using claim-based patent indicators, insights can be provided on their status of technological development.

2. Data and method

2.1. Procedure to collect the full-text sample of patent documents on nanocellulose

The overall procedure to collect the full-text document sample is shown in Fig. 1. The first step involves seeking in the bibliographic data of patent documents in Derwent Innovations Index databases. We used the following search expression, which was developed considering expert opinion [52]:

> TS=(“bacterial cellulos*” OR “cellulos* crystal*” OR “cellulos* nanocrystal*” OR “cellulos* whisker*” OR “cellulos* microcrystal*” OR “cellulos* nanowhisker*” OR “nanocrystal* cellulos*” OR “cellulos* nano-whisker*” OR “cellulos* nano-crystal*” OR “nano-crystal cellulos*” OR “cellulos* micro-crystal*” OR “cellulos* microfibril*” OR “microfibril* cellulos*” OR “cellulos* nanofibril*” OR “nanofibril* cellulos*” OR “micro-fibril* cellulos*” OR “nano-fibril* cellulos*” OR “cellulos* micro-fibril*” OR “cellulos* nano-fibril*” OR “cellulos* nanofiber*” OR nanocellulos* OR “cellulos* nanoparticle*” OR “nano-cellulos*” OR “nanopartícl* cellulos*” OR “nanosiz* cellulos*” OR “cellulos* nanofill*” OR “nano-siz* cellulos*” OR “cellulos* nano-fiber*” OR “cellulos* nano-particle*” OR “cellulos* nano-fill*” OR “nano-particl* cellulos*”)

The search was conducted in the topic field (TS), which seeks for the given terms in the title and abstract fields of the bibliographic data. The Derwent Innovations Index database was selected in this step because it enhanced the bibliographic data before indexing, especially the titles and abstracts, with information from the content of the full-text documents [53]. Thus, it improves the recall efficiency. Because we are interested in mining the full text documents, after collecting a total of 4098 bibliographic records retrieved on 28 September 2016, we separated those that contain at least one patent filed at the USPTO, resulting in 1124 bibliographic records. Next, documents were downloaded automatically from the USPTO full-text database using the GETIPDL software [54]. After converting the final set of patent documents from Hypertext Markup Language (HTML) to plain text (TXT), we imported the collected full-text documents into the text mining software Vantage Point (version 5.0) [55] and removed eventual duplicates. The focal time frame spans 1995 to 2014 - considering the earliest priority year of each patent application - and the final sample comprised 980 non-duplicated full-text patent documents.

2.2. Procedure for mining patent claims

The procedure for mining patent claims involved six steps, which were performed using the text mining software:

1) From the full-text document, which contains many parts – such as bibliographic data, background of the invention, summary of the invention, etc. – we separated the part designated “claims”;
2) After segmenting each sentence, the independent claims were separated from dependent claims [32]. In this step, we considered the fact that every dependent claim must be associated to another claim, thus we could easily tag the dependent claims using the markers presented in Table A.1. These tags were identified previously from a smaller sample of documents and then improved for the final sample analyzed;
3) All claim sentences had their preamble separated from the body of the claim tagging the phrases of transitions shown in Table A.2. Again, these tags were identified previously from a smaller sample of documents and then improved for the final sample analyzed;
4) The preambles, which are basically short phrases, were classified into one of the categories presented in Table 2. These categories were formed according to the typology available in the WIPO Patent Drafting Manual [32]. We managed to perform this task automatically due to the regular presence of words that suggest the category, such as “method”, “process”, “apparatus”, “system”, “use”, and so on;
5) The body of the claims, which can contain longer sentences than those of the preamble, were processed using the natural language processing routine from the text mining software; noun-phrases resulted from this step;
6) The final step included eliminating the stopwords and grouping similar terms (mainly singular and plural) to achieve the final set of terms from the body of the claims.

2.3. Developing indicators from patent titles and abstracts

To prove the advantage of using patent claims as a source for indicators, we compared the efficiency of classifying all extracted terms into types of nanocellulose. This comparison involved both terms extracted from claims (according to section 2.2) and terms extracted from the original titles and abstracts found in the full-text document. The idea is to evaluate whether the proposed

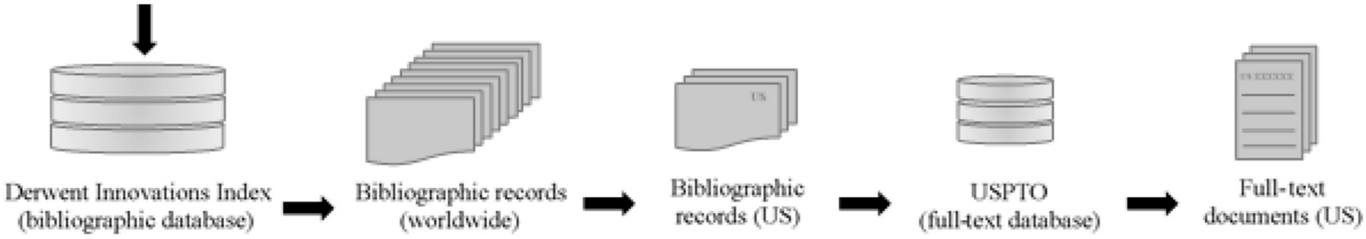

Fig. 1. Procedure to collect the full-text patent documents of nanocellulose filed in the USPTO.
Source: Developed by the authors.

Table 2
Typology and category of claims.

Typology	Category	Description
Physical entities	Apparatus, devices and systems	Claims that describe physical entities which are accessories to the main product or composition
	Compositions and products	Claims related to compositions and products that are resulted from the combination of compounds or substances, including products resulting from processes
Activities	Process and methods	Claims that characterize procedures and entities involved in processing activities, manufacturing and specific methods
	Use	Claims regarding the use of compositions and products

Source: developed by the authors based on WIPO Patent Drafting Manual [32].

method is more effective than the traditional one based only on title/abstract. To extract terms from titles and abstracts, we separated them from the other parts of the full text patent documents and applied the natural language processing routine from the text mining software. Then, we eliminated stopwords and grouped similar terms (a routine similar to the one applied to the body of patent claims).

To identify the types of nanocellulose, we develop a standardized thesaurus to categorize noun-phrases extracted from titles/abstracts and claims, as can be seen in Table A.3. Each term was carefully categorized and experts in nanocellulose were consulted to compile the final thesaurus. Generic terms or non-relevant ones were labeled “other terms”. One issue was the fact that we decided not to group noun-phrases spelled differently (for example, cellulose nanofibrils and microfibrillated cellulose, although they are the same nanomaterial), because we would lose the vocabulary used by industry. The final thesaurus was applied in noun-phrases from titles/abstracts and from claims. In the case of the claims, we considered terms from both preamble and body of the claim, both dependent and independent.

2.4. *Exploring other nanocellulose indicators from patent claims*

To explore other potential uses of patent claims in the context of technological monitoring of nanocellulose, we developed and analyzed a set of indicators. These indicators involved mainly types of nanocellulose. The general patenting activity for each type in the time span between 1995 and 2014 was obtained to get an overall idea of their development.

Another indicator is percentage ratio involving dependent claim and independent claim. The idea of this indicator is to verify if a type of nanocellulose appears more frequently in dependent claims - i.e. they are only part of the details of the invention - or in independent claims — which would mean that specific nanocellulose is part of a broad scope of the invention. Our hypothesis is a subject of matter trends to appear more in independent claims when they are considered part of the novelty.

We also developed two maps that clustered the main technological subdomains and the type of nanocellulose using the VOSviewer software (version 1.6.5) [56] algorithm. This indicator can suggest to which sector each type of nanocellulose is more associated. Furthermore, the technological subdomains are obtained by grouping patents according to their IPC code in predetermined domains and subdomains proposed by the Observatoire des Sciences et des Techniques [57].

Finally, trends of patenting were observed for each category of claims (Apparatus-Devices-Systems, Compositions-Products, Process-Methods, and Use). Our hypothesis is the percentage of claims where these categories occurs may vary and bring some insightful information.

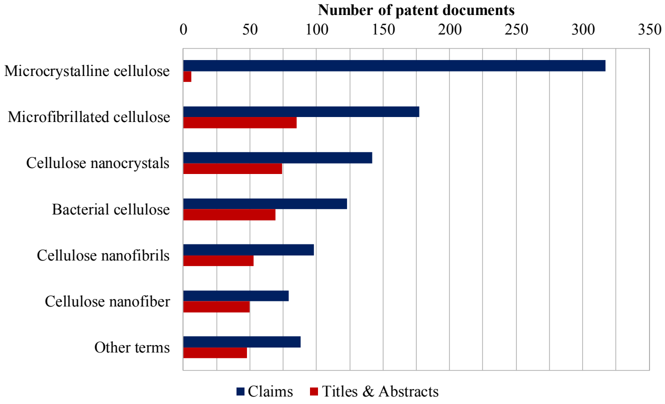

Fig. 2. Comparative efficiency of claims and titles/abstract as source of text to compile indicators.
Source: USPTO.

3. Results and discussion

3.1. Efficiency and correctness of the proposed method

Claims are a richer source of noun-phrase terms than titles and abstracts not only because they contain more words, but because the patentee must declare in detail what the invention is in this part of the document. Consequently, more documents were labeled with at least one type of nanocellulose when claims are used as the source, according to Fig. 2. The efficiency of the proposed method is 88.3% while the efficiency of categorizing terms from titles/abstracts is considerably lower (36.2%), which would affect trend analysis if only terms extract from title/abstract were used in an indicator. To avoid false positive and false negative results, we checked the title/abstract and claims of a sample of documents to confirm whether the process was correctly performed, and we verified that both routines had been carried out properly.

The higher efficiency for the claims indicators over title/abstract indicators can also be associated to the aims of these fields when a patent is being drafted. According to the WIPO Patent Drafting Manual [32], usually, the title broadly describes the invention while the abstract should define “the invention very clearly in the fewest possible words”. Furthermore, it is known that patentees usually have less interest in explaining in detail their invention, because it would facilitate the information retrieval of their patent document by a competitor, for example. Our results show that if only the title and abstract are used as the only source to develop indicators, microcrystalline cellulose may not be highlighted in technological assessment. In fact, this cellulose-based material has been greatly used as a component of pharmaceutical composition and food preparations available on the market for several years already [for instance, see Ref. 58]. However, many experts have not considered it as a cellulose in nanoscale since the nanocrystals from microcrystalline cellulose are agglomerated during the process of obtaining it – losing its “nano” effect. Moreover, some researchers have even considered microcrystalline cellulose as a source to obtain nanocrystalline cellulose, for instance by the acid hydrolysis process [44].

3.2. Claim-based patent indicators for nanocellulose

The number of dependent claims related to microcrystalline cellulose is almost five times higher than those for other cellulose-based materials, as shown in Fig. 3. Since independent claims are directly linked to the number of inventions claimed by a single patent [12], this result may also indicate that the developments involving microcrystalline celluloses are not focused on the material because the term appears more in dependent claims, i.e. detailing the components of independent claims. When independent claims contained the type of nanocellulose, it would indicate a technology related to this cellulose-based nanomaterial. We confirmed this result, checking manually some documents and consulting experts. Fig. 4 shows that microcrystalline cellulose is strongly connected to the Pharmaceutical and Cosmetic technological subdomain, confirming that this material is a component of pharmaceutical or cosmetic compositions and cannot be considered in the whole scope of nanocellulose analysis. Furthermore, there is a decrease in the number of patent documents among the periods analyzed, as seen from Fig. 5.

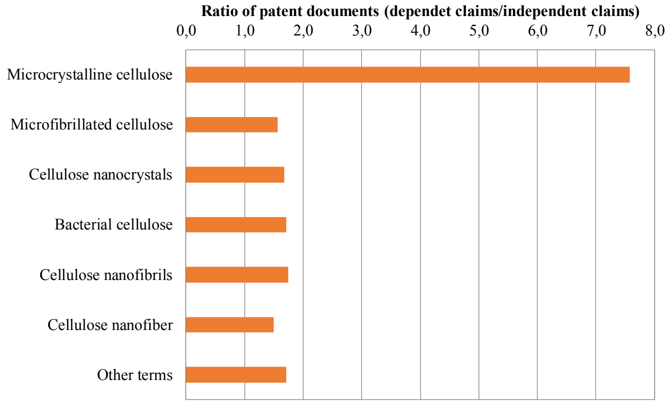

Fig. 3. Ratio between the number of patent with dependent and independent claims for the types of nanocellulose.
Source: USPTO.

For the other materials - all considered celluloses in nanoscale [44,47–49], there is a regular balance in the ratio of dependent and independent claims (1.5). Bacterial cellulose has a close connection to Medical Engineering and Biotechnology (see Fig. 4), however the number of patent documents has recently been decreasing (Fig. 5), probably because of the challenge to obtain large quantities of the nanomaterial for industrial applications through bacterial fermentation process, even though they are still relevant when high purity nanomaterial is required [46,47,59].

The other types of nanocellulose were clustered near to engineering-related technological subdomains, especially Macromolecular Chemistry (i.e. Polymers) and Materials Processing, as can be seen from Fig. 4. These strong associations may indicate that the technological development of nanocellulose is indeed dependent on initiatives in engineering applications, mainly as reinforcement in composites, or in solving technical issues, such as surface treatments of the nanofibers to improve its compatibility with polymers and other manufacturing issues. For these types of nanocellulose, patent activity has also been increasing rapidly in recent years, according to Fig. 5.

Regardless of being an independent or dependent claim, the categories of Composition-Product and Method-Process gathered a high number of patent documents – 92% and 85%, respectively. Furthermore, the Apparatus-Device-System category comprised 11% while the Use category occurred in 7.3% of patent documents. It is important to highlight that one given patent document could be categorized in more than one category as they may have different categories of claims to protect all aspects from the invention.

One interesting result emerged when we checked the percentage number of claims associated to each category in the whole sample. We observed a quick growth in the category Use in the last period analyzed, as shown in Fig. 6. We verified the content of these use-related claims and verified it could be about use of substance, product or process. For instance, one patent claimed the use of microfibrillated cellulose as an additive of concrete mixture (patent number: US20120227633); another claimed the use of a process to obtain cellulose nanofibrils with low energy expenditure (patent number: US20150158955); cellulose nanocrystals were also claimed to be a component of anti-icing and de-icing compositions in aircrafts (patent number: US20120153214). Consequently, patentees are concerned about the final application

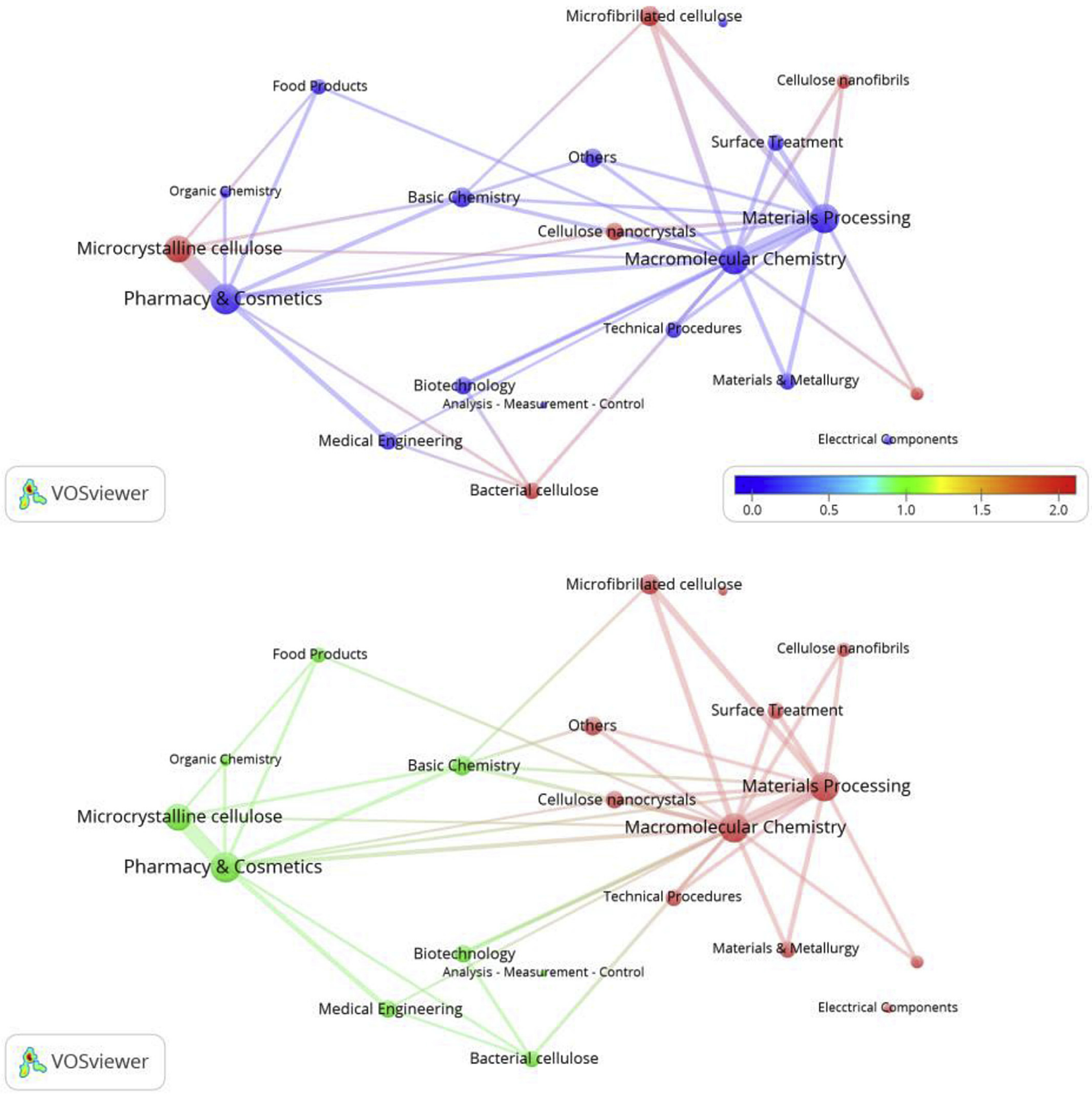

Fig. 4. Maps of networks between (a) the main technological subdomain (blue) and types of nanocellulose (red); and (b) both clustered according to the VOSviewer clustering algorithm. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
Source: USPTO.

of their invention, which indicates the potential market of their technology.

4. Conclusion

The text mining-based method proposed in this research allowed us to depict useful claim-based patent indicators to understand the detailed technical content in the field of nanocellulose, especially for the types of this nanomaterial. In other words, by using the rule of writing a patent claim, we could improve the text mining routine and compile interesting patent indicators. The results showed that patent claims are a more accurate and reliable source of terms to develop indicators than, for instance, titles and abstracts.

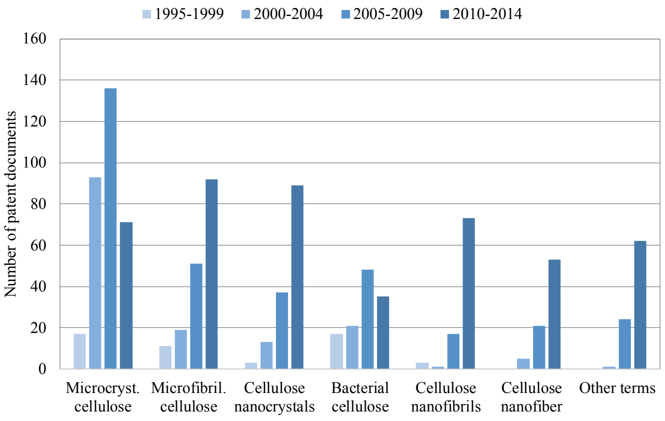

Fig. 5. Number of patent documents for the periods between 1995 and 2014 for the types of nanocellulose.
Source: USPTO.

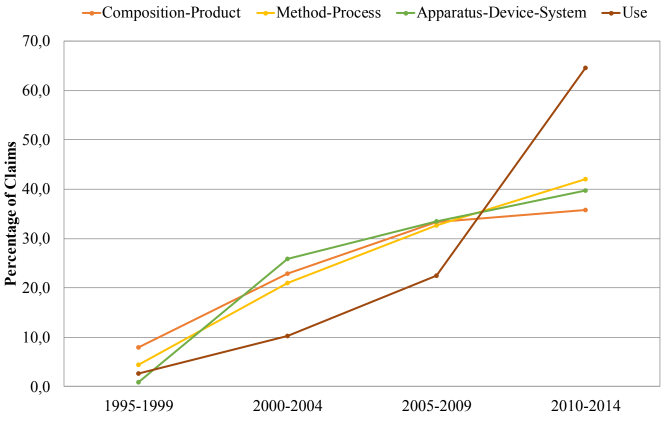

Fig. 6. Percentage of claims for each preamble category in different periods.
Source: USPTO.

We provided a discussion about this aspect in order to contribute to the development of new methods that aim to use the free texts of patent documents. The indicators assembled from patent claims in combination with other traditional indicators developed from bibliographic patent data may contribute significantly to the analytical process in monitoring, technological forecasting and competitive intelligence studies. The claim-based indicators can provide relevant trends of developments, products and manufacturing issues, applications, such as those observed for nanocellulose. Another use of the proposed method to analyze patent claims is that it can be used for assembling keywords for patent documents, a field of information normally not found in bibliographic patent data (unlike for scientific bibliographic publications that generally represent this field). These keywords obtained by text mining the patent claims might also improve, for instance, search mechanisms and information retrieving efficiency.

The method was developed considering patents filed in the United States Patents and Trademarks Office (USPTO), but the principle of using the claim structure in the text mining routine for the development of indicators can be applied to patent documents from any other offices. We also applied the analysis to nanocellulose case, but any other technological field or subject can be a target of investigation. One limitation of the proposed method is the manual checking necessary to depict tags from the patent claims and separate the preamble and the body of the claim. However, one may notice that these tags tend to be repeated over many documents, thus the labor is concentrated in the earlier application of the method. Future work will combine patent claim indicators obtained from other patent offices and other databases, compare claim indicators from granted and non-granted patents, and the influence of claims on citation analysis.

Acknowledgments

We would like to thank Dr. José Manoel Marconcini and Dr. Cauê Ribeiro Oliveira, both from the Brazilian Agricultural Research Corporation (EMBRAPA), for their expert opinion on nanocellulose. We are also thankful to the São Paulo Research Foundation (FAPESP) - process number: 12/16573-7 and 15/18878-8 - for the financial support.

Appendices A

Table A.1
Regular expression used to tag dependent claims.

<table>
<tr><td>claim</td></tr>
<tr><td>clam</td></tr>
<tr><td>as set forth in [0–9]{1,3}</td></tr>
<tr><td>according to [0–9]{1,3}</td></tr>
<tr><td>according to item [0–9]{1,3}</td></tr>
<tr><td>according [0–9]{1,3}</td></tr>
<tr><td>as claimed in [0–9]{1,3}</td></tr>
<tr><td>as claimed [0–9]{1,3}</td></tr>
<tr><td>in accordance with [0–9]{1,3}</td></tr>
<tr><td>as in [0–9]{1,3}</td></tr>
<tr><td>as defined in [0–9]{1,3}</td></tr>
<tr><td>defined by [0–9]{1,3}</td></tr>
<tr><td>defined in [0–9]{1,3}</td></tr>
<tr><td>produced by [0–9]{1,3}</td></tr>
<tr><td>prepared by [0–9]{1,3}</td></tr>
<tr><td>as set forth in [0–9]{1,3}(and | or | to |-| - | -|- | through)[0–9]{1,3}</td></tr>
<tr><td>according to [0–9]{1,3}(and | or | to |-| - | -|- | through)[0–9]{1,3}</td></tr>
<tr><td>according to items [0–9]{1,3}(and | or | to |-| - | -|- | through)[0–9]{1,3}</td></tr>
<tr><td>according [0–9]{1,3}(and | or | to |-| - | -|- | through)[0–9]{1,3}</td></tr>
<tr><td>as claimed in [0–9]{1,3}(and | or | to |-| - | -|- | through)[0–9]{1,3}</td></tr>
<tr><td>as claimed [0–9]{1,3}(and | or | to |-| - | -|- | through)[0–9]{1,3}</td></tr>
<tr><td>in accordance with [0–9]{1,3}(and | or | to |-| - | -|- | through)[0–9]{1,3}</td></tr>
<tr><td>as in [0–9]{1,3}(and | or | to |-| - | -|- | through)[0–9]{1,3}</td></tr>
<tr><td>as defined in [0–9]{1,3}(and | or | to |-| - | -|- | through)[0–9]{1,3}</td></tr>
<tr><td>defined by [0–9]{1,3}(and | or | to |-| - | -|- | through)[0–9]{1,3}</td></tr>
<tr><td>defined in [0–9]{1,3}(and | or | to |-| - | -|- | through)[0–9]{1,3}</td></tr>
<tr><td>produced by [0–9]{1,3}(and | or | to |-| - | -|- | through)[0–9]{1,3}</td></tr>
<tr><td>prepared by [0–9]{1,3}(and | or | to |-| - | -|- | through)[0–9]{1,3}</td></tr>
</table>

Source: developed by the authors.

Table A.2
Expression used to tag the transition phrases in claims.

comprising	where
comprise	in which
including	which
containing	whose
characterized by	having
characterized in that	combination of
characterized for	expressing
characterize by	capable of
consisting of	by combining
wherein	

Source: developed by the authors.

Table 3.A
Regular expression for each type of nanocellulose.

<table>
<tr><th>Types of nanocelluloses</th><th colspan="2">Regular expressions</th></tr>
<tr><td>Bacterial cellulose</td><td>.*bacterial cellulos(e|ic).*</td><td>.*gluconacetobacter xylinus cellulos(e|ic).*</td></tr>
<tr><td></td><td>.*microorganism-produced cellulose</td><td>.*cellulos(e|ic)[a-z\-]*bacterial.*</td></tr>
<tr><td></td><td>.*bacterial cellulosome.*</td><td>BC$</td></tr>
<tr><td></td><td>.*bacterial nanofiber.*</td><td>isoletade BC$</td></tr>
<tr><td></td><td>.*BNC.*</td><td>.*microbial cellulos(e|ic).*</td></tr>
<tr><td></td><td>.*biosynthetic cellulos(e|ic).*</td><td>.*microbially-derived cellulos(e|ic).*</td></tr>
<tr><td>Microfibrillated cellulose</td><td>.*cellulos(e|ic) microfibril.*</td><td>microfibril cellulose</td></tr>
<tr><td></td><td>.*microfibrillated cellulos.*</td><td>.*microfibrillar cellulos(e|ic).*</td></tr>
<tr><td></td><td>.*micro(-|)fibrillated cellulos.*</td><td>.*MFC.*</td></tr>
<tr><td></td><td>.*cellulos(e|ic) micro-fibril.*</td><td>.*cellulos(e|ic) [a-z]* [a-z]* [a-z]* microfibrillar.*</td></tr>
<tr><td></td><td>.*micro fibril cellulos(e|ic).*</td><td>.*cellulos(e|ic) fibril.*</td></tr>
<tr><td></td><td>.*microfibril cellulos(e|ic).*</td><td>.*macrofibrillated cellulos(e|ic).*</td></tr>
<tr><td>Cellulose nanofibrils</td><td>.*cellulos(e|ic) nanofibril.*</td><td>.*nano-fibrillar cellulos(e|ic).*</td></tr>
<tr><td></td><td>.*nanofibrillated cellulos.*</td><td>.*nano(-|)fibrillated cellulos.*</td></tr>
<tr><td></td><td>.*nanofibrillar cellulos(e|ic).*</td><td>.*cellulos(e|ic) nano-fibril.*</td></tr>
<tr><td>Microcrystalline cellulose</td><td>.*cellulos(e|ic) microcr.stal.*</td><td>.*micro(-|)cr.stalline cellulos(e|ic).*</td></tr>
<tr><td></td><td>.*cellulos(e|ic) micro-cr.stal.*</td><td>.*microcrist.lline cellulos(e|ic).*</td></tr>
<tr><td></td><td>.*cellulos(e|ic)-microcr.stalline.*</td><td>.*microcrystaline cellulos(e|ic).*</td></tr>
<tr><td></td><td>.*microcr.stalline cellulos(e|ic).*</td><td>.*micro(-fine |crystal | |-|fibrous)cellulos(e|ic).*</td></tr>
<tr><td>Cellulose nanofibers</td><td>.*cellulos(e|ic) nanofiber.*</td><td>.*cellulos(e|ic) nano(-|)fiber.*</td></tr>
<tr><td></td><td>.*cellulos(e|ic) nanofibre.*</td><td>.*cellulos(e|ic) nano(-|)fibre.*</td></tr>
<tr><td>Cellulose nanocrystals</td><td>.*cellulos(e|ic) nanocrystal.*</td><td>.*(NCC|CNC).*</td></tr>
<tr><td></td><td>.*nanocrystalline cellulos(e|ic).*</td><td>.*nano(|-)crystal particle/cellulos(e|ic).*</td></tr>
<tr><td></td><td>.*cellulos(e|ic) whisker.*</td><td>cellulos(e|ic) crystal.*</td></tr>
<tr><td></td><td>.*cellulos(e|ic) nanowhisker.*</td><td>crystalline cellulos(e|ic).*</td></tr>
<tr><td></td><td>.*cellulos(e|ic) nano(-|)whisker*</td><td>crystalline [a-z]* cellulos(e|ic).*</td></tr>
<tr><td></td><td>.*cellulos(e|ic) nano-crystal*</td><td>cellulos(e|ic) [a-z]* crystal.*</td></tr>
<tr><td></td><td>.*nano-crystalline cellulos(e|ic).*</td><td></td></tr>
<tr><td>Other terms</td><td>.*microfine cellulos(e|ic).*</td><td>.*nanoparticle cellulos(e|ic).*</td></tr>
<tr><td></td><td>.*cellulos(e|ic) microfiber.*</td><td>.*cellulos(e|ic) nano-particle.*</td></tr>
<tr><td></td><td>.*cellulose microtubule.*</td><td>.*nano-particl* cellulos(e|ic).*</td></tr>
<tr><td></td><td>.*cellulos(e|ic) nanofilament.*</td><td>.*nanosized cellulos(e|ic).*</td></tr>
<tr><td></td><td>.*cellulos(e|ic) nanotubule.*</td><td>.*nano-sized cellulos.*</td></tr>
<tr><td></td><td>.*cellulos(e|ic) nanometer.*</td><td>.*nanosized [a-z]* [a-z]* cellulos(e|ic).*</td></tr>
<tr><td></td><td>.*nano-dispersed cellulos(e|ic).*</td><td>.*cellulos(e|ic) nanofiller.*</td></tr>
<tr><td></td><td>.*nanocellulos(e|ic).*</td><td>.*cellulos(e|ic) nano-filler.*</td></tr>
<tr><td></td><td>.*nano-cellulos(e|ic).*</td><td>.*algal cellulose.*</td></tr>
<tr><td></td><td>.*cellulos(e|ic) nanoparticle.*</td><td>.*cyanobacterium [a-z].* cellulos(e|ic).*</td></tr>
<tr><td></td><td>.*nanoparticled cellulos(e|ic).*</td><td>.*tunicate cellulos(e|ic).*</td></tr>
</table>

Source: developed by the authors.

References

[1] K. Pavitt, Patent statistics as indicators of innovative activities: possibilities and problems, Scientometrics 7 (1985) 77–99, http://dx.doi.org/10.1007/BF02020142.

[2] A. Abbas, L. Zhang, S.U. Khan, A literature review on the state-of-the-art in patent analysis, World Pat. Inf. 37 (2014) 3–13, http://dx.doi.org/10.1016/j.wpi.2013.12.006.

[3] A.F.J. van Raan, Patent citations analysis and its value in research evaluation: a review and a new approach to map technology-relevant research, J. Data Inf. Sci. 2 (2017) 13–50, http://dx.doi.org/10.1515/jdis-2017-0002.

[4] F. Madani, C. Weber, The evolution of patent mining: applying bibliometrics analysis and keyword network analysis, World Pat. Inf. 46 (2016) 32–48, http://dx.doi.org/10.1016/j.wpi.2016.05.008.

[5] M.E. Mogee, Patents and technology intelligence, in: Keep. Abreast Sci. Technol. Tech. Intell. Bus., Battelle Press, Columbus, 1997.

[6] D. Rotolo, I. Rafols, M. Hopkins, L. Leydesdorff, Scientometric Mapping as a Strategic Intelligence Tool for the Governance of Emerging Technologies, 2014.

[7] M. Fattori, G. Pedrazzi, R. Turra, Text mining applied to patent mapping: a practical business case, World Pat. Inf. 25 (2003) 335–342, http://dx.doi.org/10.1016/S0172-2190(03)00113-3.

[8] A.F. Breitzman, Assessing an industry's R&D focus rapidly: a case study using data-driven categorization in a consumer products area, Compet. Intell. Rev. 11 (2000) 58–64.

[9] J.P. Courtial, M. Callon, A. Sigogneau, The use of patent titles for identifying the topics of invention and forecasting trends, Scientometrics 26 (1993) 231–242.

[10] N. van Zeebroeck, B. van Pottelsberghe de la Potterie, The vulnerability of patent value determinants, Econ. Innov. New Technol. 20 (2011) 283–308, http://dx.doi.org/10.1080/10438591003668638.

[11] F. Caviggioli, Technology fusion: identification and analysis of the drivers of technology convergence using patent data, Technovation 55–56 (2016) 22–32, http://dx.doi.org/10.1016/j.technovation.2016.04.003.

[12] G. Kim, J. Bae, A novel approach to forecast promising technology through patent analysis, Technol. Forecast. Soc. Change 117 (2017) 228–237, http://dx.doi.org/10.1016/j.techfore.2016.11.023.

[13] J.O. Lanjouw, M. Schankerman, Patent quality and research productivity: measuring innovation with multiple indicators*, Econ. J. 114 (2004) 441–465, http://dx.doi.org/10.1111/j.1468-0297.2004.00216.x.

[14] L.F. Chanchetti, S.M. Oviedo Diaz, D.H. Milanez, D.R. Leiva, L.I.L. de Faria, T.T. Ishikawa, Technological forecasting of hydrogen storage materials using patent indicators, Int. J. Hydrogen Energy 41 (2016) 18301–18310, http://dx.doi.org/10.1016/j.ijhydene.2016.08.137.

[15] Organisation for Economic Co-operation and Development, OECD Patent Statistics Manual, 2009, http://dx.doi.org/10.1787/9789264056442-en.

[16] R.J. Tijssen, Global and domestic utilization of industrial relevant science: patent citation analysis of science–technology interactions and knowledge flows, Res. Policy 30 (2001) 35–54, http://dx.doi.org/10.1016/S0048-7333(99)00080-3.

[17] Y.-H. Tseng, C.-J. Lin, Y.-I. Lin, Text mining techniques for patent analysis, Inf. Process. Manag. 43 (2007) 1216–1247, http://dx.doi.org/10.1016/j.ipm.2006.11.011.

[18] D.H. Milanez, R.M. do Amaral, L.I.L. de Faria, J.A.R. Gregolin, Technological indicators of nanocellulose advances obtained from data and text mining applied to patent documents, Mater. Res. 17 (2014) 1513–1522, http://dx.doi.org/10.1590/1516-1439.266314.

[19] S. Sheremetyeva, Natural language analysis of patent claims, in: Proc. ACL-2003 Work. Pat. Corpus Process, Association for Computational Linguistics, Morristown, NJ, USA, 2003, pp. 66–73, http://dx.doi.org/10.3115/1119303.1119311.

[20] L. Wanner, R. Baeza-Yates, S. Brügmann, J. Codina, B. Diallo, E. Escorsa, M. Giereth, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, G. Piella, I. Puhlmann, G. Rao, M. Rotard, P. Schoester, L. Serafini, V. Zervaki, Towards content-oriented patent document processing, World Pat. Inf. 30 (2008) 21–33, http://dx.doi.org/10.1016/j.wpi.2007.03.008.

[21] Y. Xu, Apply text mining in analysis of patent document, in: 2009 IEEE 10th Int. Conf. Comput. Ind. Des. Concept. Des., 2009, pp. 2350–2352, http://dx.doi.org/10.1109/CAIDCD.2009.5375302.

[22] S.-Y. Yang, V.-W. Soo, Extract conceptual graphs from plain texts in patent claims, Eng. Appl. Artif. Intell. 25 (2012) 874–887, http://dx.doi.org/10.1016/j.engappai.2011.11.006.
[23] B. Yoon, Y. Park, A systematic approach for identifying technology opportunities: keyword-based morphology analysis, Technol. Forecast. Soc. Change 72 (2005) 145–160, http://dx.doi.org/10.1016/j.techfore.2004.08.011.
[24] J. Yoon, K. Kim, Detecting signals of new technological opportunities using semantic patent analysis and outlier detection, Scientometrics 90 (2011) 445–461, http://dx.doi.org/10.1007/s11192-011-0543-2.
[25] J. Yoon, K. Kim, An analysis of property–function based patent networks for strategic R&D planning in fast-moving industries: the case of silicon-based thin film solar cells, Expert Syst. Appl. 39 (2012) 7709–7717, http://dx.doi.org/10.1016/j.eswa.2012.01.035.
[26] J. Yoon, H. Park, K. Kim, Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis, Scientometrics 94 (2012) 313–331, http://dx.doi.org/10.1007/s11192-012-0830-6.
[27] A. Porter, D. Chiavetta, Introduction to special issue on TechMining, Scientometrics 100 (2014) 611–612, http://dx.doi.org/10.1007/s11192-014-1340-5.
[28] S. Verberne, E. D'hondt, N. Oostdijk, Quantifying the challenges in parsing patent claims, in: 1st Int. Work. Adv. Pat. Inf. Retr., Milton Keynes, UK, 2010 doi:970.
[29] B. Yoon, Y. Park, A text-mining-based patent network: analytical tool for high-technology trend, J. High Technol. Manag. Res. 15 (2004) 37–50, http://dx.doi.org/10.1016/j.hitech.2003.09.003.
[30] L. Zhang, D. Zhu, Research of technical development trend and hot points based on text mining, in: 2010 2nd Int. Conf. Inf. Eng. Comput. Sci., 2010, pp. 1–5, http://dx.doi.org/10.1109/ICIECS.2010.5678397.
[31] USPTO, Nonprovisional (Utility) Patent Application Filing Guide, 2009. https://www.uspto.gov/patents-getting-started/patent-basics/types-patent-applications/nonprovisional-utility-patent#heading-18. (Accessed 1 August 2017).
[32] W.I.P.O. WIPO, WIPO Patent Drafting Manual, 2010, p. 138. www.wipo.int/freepublications/en/patents/867/wipo_pub_867.pdf. (Accessed 5 September 2013).
[33] D.H. Lin, S.C. Hsieh, Characteristics of independent claim: a corpus-linguistic approach to contemporary english patents, Comput. Linguist. Chin. Lang. Process. 16 (2011) 77–106.
[34] H. Mase, T. Matsubayashi, Y. Ogawa, M. Iwayama, T. Oshio, Proposal of two-stage patent retrieval method considering the claim structure, ACM Trans. Asian Lang. Inf. Process. 4 (2005) 190–206, http://dx.doi.org/10.1145/1105696.1105702.
[35] J. Shin, Y. Park, Generation and application of patent claim map: text mining and network analysis, J. Intellect. Prop. Rights 10 (2005) 198–205.
[36] A. Fujii, M. Iwayama, N. Kando, Introduction to the special issue on patent processing, Inf. Process. Manag. 43 (2007) 1149–1153, http://dx.doi.org/10.1016/j.ipm.2006.11.004.
[37] C.H. Hong, D.S. Han, Nylon-4 Composite, US20110086948, 2009.
[38] D. Zivković, M. Niculović, D. Manasijević, D. Minic, V. Cosovic, M. Sibinović, Bibliometric trend and patent analysis in nano-alloys research for period 2000-2013, (n.d.).
[39] G. Hu, W. Liu, Nano/micro-electro mechanical systems: a patent view, J. Nanoparticle Res. 17 (2015) 465, http://dx.doi.org/10.1007/s11051-015-3273-1.
[40] D. Neuman, J.N. Chandhok, Patent watch: nanomedicine patents highlight importance of production methods, Nat. Rev. Drug Discov. 15 (2016) 448–449, http://dx.doi.org/10.1038/nrd.2016.118.
[41] S.K. Arora, A.L. Porter, J. Youtie, P. Shapira, Capturing new developments in an emerging technology: an updated search strategy for identifying nanotechnology research outputs, Scientometrics 95 (2012) 351–370, http://dx.doi.org/10.1007/s11192-012-0903-6.
[42] M. Igami, T. Okazaki, Capturing Nanotechnology's Current State of Development via Analysis of Patents, Paris, 2007, http://eric.ed.gov/ERICWebPortal/recordDetail?accno=ED504018. (Accessed 4 July 2012).
[43] C. Huang, A. Notten, N. Rasters, Nanoscience and technology publications and patents: a review of social science studies and search strategies, J. Technol. Transf. 36 (2011) 145–172, http://dx.doi.org/10.1007/s10961-009-9149-8.
[44] S.J. Eichhorn, A. Dufresne, M. Aranguren, N.E. Marcovich, J.R. Capadona, S.J. Rowan, C. Weder, W. Thielemans, M. Roman, S. Renneckar, W. Gindl, S. Veigel, J. Keckes, H. Yano, K. Abe, M. Nogi, a.N. Nakagaito, A. Mangalam, J. Simonsen, a.S. Benight, A. Bismarck, L.a. Berglund, T. Peijs, Review: current international research into cellulose nanofibres and nanocomposites, J. Mater. Sci. 45 (2010) 1–33, http://dx.doi.org/10.1007/s10853-009-3874-0.
[45] Research and Markets: The Global Market for Nanocellulose to 2017, Updated 2013 Report, Fort Mill Times, Fort Mill, SC, (n.d.). http://www.fortmilltimes.com/2013/04/29/2651687/research-and-markets-the-global.html (Accessed 29 April 2013).
[46] A. Dufresne, Nanocellulose: a new ageless bionanomaterial, Mater. Today 16 (2013) 220–227.
[47] D. Klemm, F. Kramer, S. Moritz, T. Lindström, M. Ankerfors, D. Gray, A. Dorris, Nanocelluloses: a new family of nature-based materials, Angew. Chem. Int. Ed. Engl. 50 (2011) 5438–5466, http://dx.doi.org/10.1002/anie.201001273.
[48] R.J. Moon, A. Martini, J. Nairn, J. Simonsen, J. Youngblood, Cellulose nanomaterials review: structure, properties and nanocomposites, Chem. Soc. Rev. 40 (2011) 3941–3994, http://dx.doi.org/10.1039/c0cs00108b.
[49] G. Siqueira, J. Bras, A. Dufresne, Cellulosic bionanocomposites: a review of preparation, properties and applications, Polym. (Basel) 2 (2010) 728–765, http://dx.doi.org/10.3390/polym2040728.
[50] H. Charreau, M.L. Foresti, A. Vazquez, Nanocellulose patents trends: a comprehensive review on patents on cellulose nanocrystals, microfibrillated and bacterial cellulose, Recent Pat. Nanotechnol. 7 (2013) 56–80. http://www.ncbi.nlm.nih.gov/pubmed/22747719.
[51] D.H. Milanez, R.M. Do Amaral, L.I.L. De Faria, J.A.R. Gregolin, Assessing nanocellulose developments using science and technology indicators, Mater. Res. 16 (2013) 635–641, http://dx.doi.org/10.1590/S1516-14392013005000033.
[52] D.H. Milanez, A.C.A. Conserva, R.M. Amaral, L.I.L. Faria, J.A.R. Gregolin, A. Carlos, A.C.A. Conserva, Análise de bases de dados e termos de busca para estudos bibliométricos e monitoramento científico em nanocelulose, Em Questão 20 (2014) 114–133.
[53] T. Scientific, Derwent Innovations Index: Tools of the Trade, first ed., Thomson Scientific, London, 2003.
[54] K. Ujihara, GetIPDL: Patent Downloader, 2010. http://www.getipdl.net/en/. (Accessed 3 March 2014).
[55] S. Technology, The VantagePoint, 2017. https://www.thevantagepoint.com/. (Accessed 3 March 2014).
[56] CWTS, VOSviewer, 2017. http://www.vosviewer.com/. (Accessed 24 January 2017).
[57] OST, Science & Technologie Indicateurs, Paris, 2010.
[58] FMC, Avicel® PH Microcrystalline Cellulose - Datasheet, 2014, p. 1. http://www.fmcbiopolymer.com/portals/pharm/content/docs/avicelphmsds.pdf. (Accessed 10 December 2014).
[59] R.J. Mooney, U.Y. Nahm, Text mining with information extraction, in: W. Daelemans, T. du Plessis, C. Snyman, L. Teck (Eds.), Multiling. Electron. Lang. Manag. Proc. 4th Int. MIDO Colloq., South Africa, 2005, pp. 141–160.

Douglas Henrique Milanez is researcher from the Centre for Technological Information in Materials and from the Centre for Technological and Organizational Intelligence Research, both from the Federal University of São Carlos (UFSCar), Brazil. He is doctor in Materials Science and Engineering (UFSCar) and he has been researching new methodologies to mining free text from patent documents in order to develop and analyze Science and Technological (S&T) Indicators in context of technological monitoring, technological forecasting and competitive intelligence. He is also interested on bibliometrics, science mapping, classification systems, S&T evaluation, and innovation management. In 2015, he was temporary professor in the Science Information Department from UFSCar, lecturing in Information Units Management and innovation Management.

Leandro Innocentini Lopes de Faria is Professor in the Information Science Department, researcher and coordinator from the Centre for Technological Information in Materials and from the Centre for Technological and Organizational Intelligence Research, both from UFSCar. He is doctor in Materials Science and Engineering at UFSCar and doctor in Information and Communication Science at Universite d'Aix-Marseille III. His research interests rely on bibliometrics, text mining techniques, S&T indicators, science mapping, technological forecasting and competitive intelligence.

Roniberto Morato do Amaral is Professor in the Information Science Department, researcher from the Centre for Technological Information in Materials and from the Centre for Technological and Organizational Intelligence Research, both from UFSCar. He is doctor in Production Engineering and Bachelor in Library and Information Science, both at UFSCar. His research interests rely on management, bibliometrics, S&T indicators, analysis of competence, technological forecasting and competitive intelligence.

José Angelo Rodrigues Gregolin is retired Professor from the Materials Engineering Department, and researcher from the Centre for Technological Information in Materials at UFSCar. He is doctor in Mechanical Engineering at the Campinas State University. His research interests rely on innovation management, bibliometrics, S&T indicators, technological forecasting and competitive intelligence.