Renewable Energy 123 (2018) 407–416

Contents lists available at ScienceDirect

Renewable Energy

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

A bibliometric-based technique to identify emerging photovoltaic technologies in a comparative assessment with expert review

Alberto Moro*, Elisa Boelman, Geraldine Joanny, Juan Lopez Garcia

European Commission, Joint Research Centre, via Enrico Fermi 2749, Ispra, Italy

ARTICLE INFO

Article history:
Available online 5 February 2018

Keywords:
Bibliometrics
Technology mining
Emerging technologies
Photovoltaics
Horizon scanning
Policy support

ABSTRACT

This paper compares the results of technology mapping from bibliometric analysis and results from expert review to identify emerging solar photovoltaic (PV) technologies. The bibliometric analysis is based on “Tools for Innovation Monitoring” (TIM), a new software code developed by the Joint Research Centre. With this text-mining software a set of relevant keywords is extracted through frequency analysis from a corpus of pertinent scientific publications. Keywords obtained by quantitative analysis by TIM are tested against results from qualitative cognitive analysis by an international panel of PV technology experts by means of a set of proposed indicators. The technologies identified by the PV experts are well represented amongst the most frequently occurring (highest ranked) keywords retrieved by TIM. The more salient keywords tend to correspond to the relatively more established technologies such as dye sensitised solar cells, organic PV and more recently-developed technologies such as perovskites. These high rated/developed keywords/technologies can be relatively straightforwardly detected through bibliometric analysis. Contrary to that, keywords designating the most emerging technologies like ferroelectric PV, hot carriers and multiple exciton generation solar cells tend to occur much less frequently and therefore provide weaker signals. These weak signals can be important in foresight.

1. Introduction

1.1. Relevance, scope and structure of this paper

This paper compares results from two different horizon scanning methods. A qualitative cognitive expert review is compared with a quantitative bibliometric analysis of keywords; these detect and monitor promising emerging technologies at an early stage of development. A case study focusing on solar photovoltaics illustrates the approach.

From energy and innovation policy perspectives, new clean energy technologies have a potential to reduce greenhouse gas emissions and to spur job creation and economic growth [1]. From a data-driven economy perspective, many EU policy domains are faced with the challenge of extracting accurate, targeted and timely information from an increasing volume of textual data. In this context, policy makers would benefit from text mining and analysis solutions to access the right information, in the proper format for the decision making process in a variety of contexts.

Within the Joint Research Centre (JRC) of the European Commission, work is on-going on bibliometric analysis as a complement to expert consultation for the detection and monitoring of emerging technologies [2,3]. This paper outlines the methodology and results of a pilot exercise on a quantitative method exploiting computer-based keyword mapping of emerging PV technologies.

In the current literature it is commonly recognised that focused expert reviews are more suitable than text mining methods in identifying weak signals [4], so results from the JRC bibliometric software are tested against the results of an expert review exercise [5] by means of a set of indicators specially conceived for this purpose. The quality, background and impartiality of the panel of elicited experts make their outputs particularly suitable to be considered as a reference.

For the purposes of this study, emerging photovoltaic (PV) technologies are broadly deemed to include novel and potentially transformative PV materials and/or production processes that are still far away from commercial deployment. Referring to the Technology Readiness Level (TRL) definition, adopted also by the European Commission [6], the present study focused on emerging

* Corresponding author.
E-mail addresses: alberto.moro@ec.europa.eu (A. Moro), elisa.boelman@ec.europa.eu (E. Boelman), geraldine.joanny@ec.europa.eu (G. Joanny), juan.lopez-garcia@ec.europa.eu (J.L. Garcia).

https://doi.org/10.1016/j.renene.2018.02.016

PV technologies with a TRL not exceeding TRL 4.

This paper is structured as follows, section 1.2 below provides a brief context on the "Tools for Innovation Monitoring" (TIM), the new software developed at the JRC; section 2 describes the general methodology adopted to compare the performance of the software-based bibliometrics approach (quantitative keyword frequency analysis) in identifying emerging technologies with the "classic" method of expert review (qualitative cognitive analysis); this comparison is implemented by means of specific indicators described in 2.3; section 3 presents results as keyword lists generated by TIM, which section 4 benchmarks and section 5 discusses in terms of keyword frequency analysis and expert knowledge.

1.2. Software-based Tools for Innovation Monitoring (TIM)

Information on scientific and patent production can complement expert knowledge by providing quantitative evidence to inform policies that are subject to an increasing integration between research and development (R&D) and technology innovation [7,8]. Expanded use of databases and the enhanced computing power nowadays allow combining bibliometric analysis, counting activity levels and identifying patterns in R&D bibliographic records plus patent analyses [9,10], with text mining from complex databases to identify, select and visualise information on emerging technologies. This allows to construct maps of keywords and R&D actors (for example: [2,3,11,12,13]), based on co-publication and co-patenting, as a valuable evidence-base to compare the advancement of knowledge amongst different technologies in the course of time and across different geographical regions.

JRC has developed a monitoring system for tracking the evolution of established and emerging technologies named Tools for Innovation Monitoring (TIM). It is based on semantic analysis, powerful data mining and visualization of complex data sets. TIM counts activity levels (based e.g. on R&D bibliography and patents) and identifies patterns of collaboration and technological evolution, potentially tracking the progression of keywords over time and by domain. TIM uses network analysis to detect events related to technology change, by identifying, clustering and visualizing complex relationships and connections by topics, institutions and countries or regions [2].

2. Methodology

Expert reviews are well established in horizon scanning and considered one of the best methods to identify weak signals [4].

The expert review of this case-study was performed in December 2016 by a good number (~20) of international PV experts with a good mix of expertise in various PV technologies. Its findings were used to benchmark the ability of TIM to retrieve keywords related to emerging PV technologies (Fig. 1). The experts analysed the degree of development, challenges and potential of the emerging PV technologies identified as relevant, together with their Technology Readiness Levels [5]. The "quantitative term frequency analysis" process (right side of Fig. 1) can be summarised as follows. First, energy analysts (different from the PV experts involved in the expert review) defined a Boolean search string (Table 2 in 2.2.2) and inserted it to the TIM Editor software (section 1.2) to retrieve a set of scientific publications relevant to new and innovative Emerging Technologies in the PV sector. From this dataset, TIM then extracted and refined a list of keywords, which was subsequently compared to the list of emerging PV technologies identified by the experts (Table 1). The quantitative comparison was performed by specifically designed indicators described in section 2.3.

2.1. Qualitative cognitive analysis by expert review

A qualitative cognitive review exercise involving internationally recognised experts on PV technologies was conducted according to common technology foresight methodologies [14]. Experts were asked to produce a list of PV technologies which can be considered "emerging technologies" as defined in section 1.1. One panel consisted of JRC in-house senior experts on PV technologies, who drew on their own knowledge and experience to identify about 10 PV emerging technologies. A second list, of about 20 PV emerging technologies, was independently proposed by another international panel of 15 experts gathered for a workshop in December

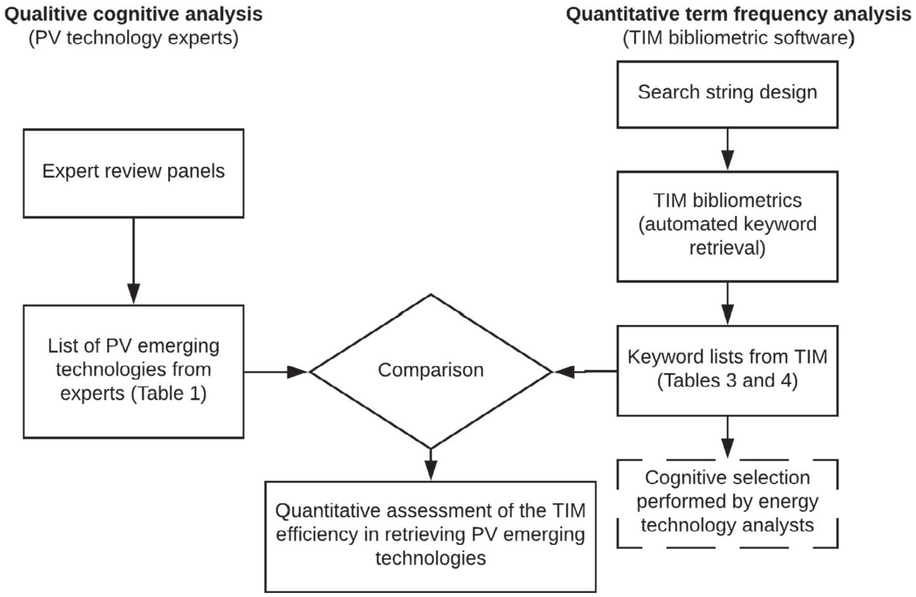

Fig. 1. Visual representation of the protocol approach adopted in this paper.

Table 1
List of emerging technologies in the PV sector identified by the expert review method.

Emerging PV technologies identified by experts
Kesterite thin film solar cells (or CZTS)
Perovskite thin film solar cells
Organic solar cells (OPV)
Dye-Sensitized Solar cells (DSSC)
Intermediate band solar cells (IBSC)
Solar cells with nanostructures
Quantum dots solar cells
nanowire solar cells
graphene and fullerene for PV applications
Plasmonic solar cells
Low-cost manufacturing processes such as roll-to-roll
flexible solar cells
Innovative multi-junction solar cells (also "multi junction")
Silicon-based tandem cells
Thermo-photovoltaics (or Thermal)
Innovative III-V compounds based solar cells (search for "III″)
Photoelectrocatalytic devices (also "photocatalytic")
Ferroelectric PV
Multiple exciton generation (MEG) solar cells
Hot carrier solar cells
Transparent conducting materials
Carrier-selective contacts
Solar cells from semiconductor foils
New photovoltaic materials via combinatorial and computational design (also: "modelling")

2016 [5]. The two lists (mainly overlapping) were then manually merged by the energy analysts (Table 1) and used to benchmark keywords retrieved from bibliometric analysis by checking to what extent the keyword lists generated by the TIM software match with the keywords underlined in Table 1 (see).

2.2. Quantitative keyword frequency analysis by TIM bibliometrics

TIM can retrieve bibliometric data from several sources such as the SCOPUS database of peer-reviewed scientific journals [15]. CORDIS, the database of European Union (EU) research projects [16], and PATSTAT, a wide database of patents [17]. For the purposes of this paper, dictionary creation and keyword extraction are the most relevant computational linguistic operations performed by TIM.

2.2.1. Dictionary creation

TIM creates a dictionary of concepts (Fig. 2) and their synonyms from a reference corpus (the whole data set analysed) of documents present in SCOPUS, CORDIS and PATSTAT published, between 1996 and 2016, in all scientific fields.

TIM extracts single- and multi-words as well as acronyms from titles, abstracts and keyword fields in the reference corpus, normalises the words (grouping instances of the same term, removing inconsistencies in e.g. spelling or word choice, ranks them by relevance and stores them as concepts in the dictionary. The terms in the dictionary are prioritised and weighted according to the processing implemented by the extractors. Among others, composite term frequency–inverse document frequency (tf-idf) weighting [18] allows ranking terms according to their number of occurrences in a document, offset by the number of occurrences in the whole corpus. TIM uses the dictionary as a central data structure when extracting “clean keywords” (see section 2.2.3) from more specific sets of documents.

2.2.2. Search string design

The dataset extracted by TIM was defined using a Boolean search string (Table 2) designed to capture future, emerging and other innovative or exploratory aspects of the PV technology, as sketched in Fig. 3 below.

The search string was designed by way of a literature review based on existing PV technology delineations from scientometrics ([8,10,19,20], and technology assessment literature [21,22]).

This string broadly delineates PV technology at a general level, with the aim of maximising the results (bibliometric recall). In addition to this technology delineation (left part of Fig. 3), the search string also includes a future/emerging attribution (right part of Fig. 3), aimed at retrieving publications with an explicit element of novelty. In order to improve retrieval quality (bibliometric precision), a proximity-search limit of up to 10 words was established between the PV-technology-delineation and the “future/emerging” attribution parts of the Boolean search string.

The bibliometric search was performed on 03.03.2017 by using the TIM Beta 2016 version. A relevant corpus of 6481 documents were identified, consisting of 131 EU projects, 717 Patents and, from the SCOPUS data base: 259 Reviews, 2164 conference proceedings, 73 book chapters and 3173 articles.

2.2.3. Keyword extraction and automated cleaning/clumping

For the purpose of identifying technologies that could be considered as emerging PV technologies, the analysis focuses on the keywords associated to the documents retrieved by TIM. In first instance, TIM lists unprocessed keywords defined by document authors or journal editors, which are referred to as “native keywords” in this paper. Secondly, TIM can also group similar words into concepts which are then listed as semantically “clean keywords”, as explained below; this process is also known in literature as “clumping” [13].

The native keyword extraction process starts by designing a Boolean search string (see Table 2 and Fig. 3) to retrieve a dataset of publications (Fig. 4) relevant to both PV and future/emerging, (years 1996–2016). TIM then extracts the native keywords associated to each publication in the dataset and calculates how many times the same keyword is used in different publications in the whole dataset. Then, it ranks-orders these keywords from the most to the least frequently occurring ones.

TIM extracted about 9800 native keywords from the whole dataset of 6481 documents identified by the search string designed as above. These keywords contain many inflected versions of the same word (for example *Solar-Cell*, *solar cells*, *solar cell* …), since they are retrieved directly as provided by the different authors/editors, During the clumping process the keywords are first stemmed (reduced to the word root) and compared to the existing terms

Table 2
Search string used by the TIM software.

ti_abs_key:(("photovoltaic future"~10 OR ″photovoltaic emerging"~10 OR ″photovoltaic innovative"~10 OR ″photovoltaic disruptive"~10 OR ″photovoltaic visionary"~10 OR ″photovoltaic exploratory"~10 OR ″photovoltaic unexpected"~10 OR ″photovoltaic new"~10 OR ″photovoltaic novel"~10) OR ("solar cell" AND (future OR emerging OR innovative OR disruptive OR visionary OR exploratory OR unexpected OR ("solar new"~10 AND ″cell new"~10) OR ("solar novel"~10 AND ″cell novel"~10))) OR ("solar PV″ AND (future OR emerging OR innovative OR disruptive OR visionary OR exploratory OR unexpected OR ("solar new"~10 AND ″PV new"~10) OR ("solar novel"~10 AND ″PV novel"~10))) NOT emergency)

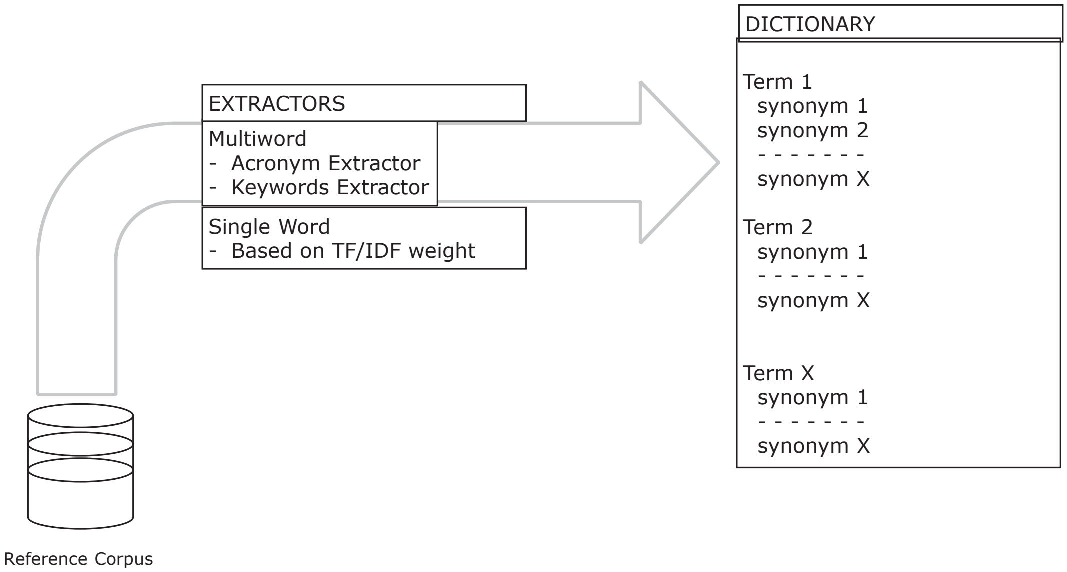

Fig. 2. Visual representation of the dictionary creation by the TIM bibliometric software.

in the dictionary of concepts [23–25]. They are then weighted (according to their tf-idf and other parameters) and rank-ordered per document in a “bag of concepts”, which is the basis for the list of clean keywords for the whole dataset. Like the native keywords, these clean keywords are rank-ordered from highest to lowest frequency of occurrence. The clean keyword list also records the frequency of occurrence of each keyword. This process resulted in a list of about 5800 clean keywords.

2.2.4. Cognitive selection by energy analysts

Keywords defined by authors or editors provide concise indications of the most important core concepts in a document, and therefore can potentially convey useful information for monitoring research trends and their evolution [26]. Rank-ordering of keywords according to their frequency of occurrence can provide an overall picture of most relevant keywords (“master keywords”) on top of a rank of less frequently occurring keywords. In terms of tracking research evolution, a bibliometric analyst can relate this notion of most frequently occurring keywords to the idea that some keywords may be used by authors to frame their own attempt at transforming a field [27]. Furthermore, analysis of the most-frequently occurring keywords can provide indications of research hot spots within given time periods. The frequency of keywords and their ranks is known to follow a power-law distribution [28].

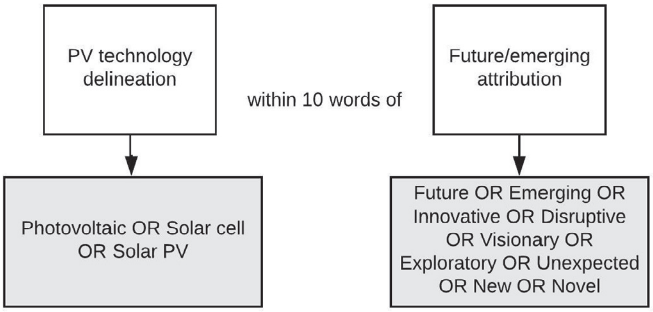
PV technology delineation
within 10 words of
Future/emerging attribution
Photovoltaic OR Solar cell OR Solar PV
Future OR Emerging OR Innovative OR Disruptive OR Visionary OR Exploratory OR Unexpected OR New OR Novel

Fig. 3. Schematics of the Boolean search string designed to catch the future/emerging aspects of PV technology.

Since the corpus of documents searched by the bibliometric software is extremely wide and not specifically devoted to emerging technologies, not necessarily the most high-ranked keywords are significant to the purpose of identifying emerging/future technologies. As can be seen from Tables 3 and Table 4, the lists of retrieved native keywords and clean keywords (after the clumping process) embed several trivial terms like “photovoltaics” and “solar cell”, or refer to mature technologies such as “silicon”, or too vague concepts like “thin film”.

In order to further refine these lists a “cognitive selection” is necessary. This is typically [13] performed by experts, or energy analysts.

TIM orders keywords by frequency (Tables 3 and 4), which the energy analysts semantically examined starting from the most frequently retrieved keywords. This was performed on the basis of own experience plus definitions and descriptions from relevant bibliographic sources. This final step is represented, in Fig. 1, by the dashed-line box “Cognitive selection of candidate PV”.

The screening work required of analysts is quite committing and cannot be realistically done for all the keywords delivered by the bibliometric software (5800 in the clean keyword list illustrated in Table 4). However, screening only a subset of the keywords normally entails loss of information. In order to decide at which rank to stop the process, there must be a trade-off between available resources (time of experts/analysts) and retrieval performance required.

From our experience, 300 is a reasonable number of keywords that an average-experienced analyst can screen in a week of work, carefully discerning one by one if they are just terms of general use or if they can represent a concept useful to designate or identify an early-stage or emerging technology. This number is confirmed by a frequency-rank analysis (section 3.2.1) and by the literature: “*The Term Clumping process … has been completed and we have obtained a review by … experts of the Top 300 consolidated terms.*” [13].

In the following section we propose some specifically designed indicators to measure the effectiveness of the bibliometric software, tested against the results provided by the expert review method.

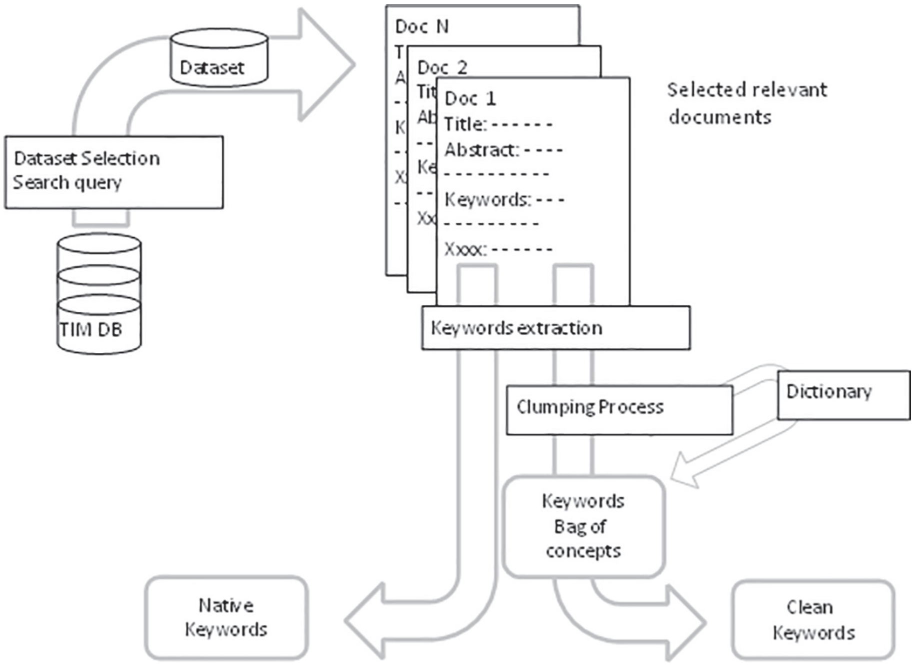

Fig. 4. Clean keyword extraction by TIM.

2.3. Comparison of findings from expert review and TIM bibliometrics

In this section we detail indicators and parameters we deem useful to benchmark or compare findings from a bibliometric software (TIM in our case) against expert reviews. Section 4 provides application examples.

The *test set* [29] used for assessing the performance of the software in retrieving specific concepts is the set of 24 underlined keywords in Table 1 we also defined "*Marker Keywords*" (MK), designating emerging PV technologies proposed by experts.

We define as “first N ranked” (N) the set of the first (or top) N most frequently occurring keywords retrieved by TIM under specified search/filtering conditions. In the numerical examples presented in section 4, considering the proposed use of the bibliometric software and our experience (see section 2.2.4) we mainly adopted N = 300, although the use of N can be generalised.

Table 3
Top most frequently occurring native keywords.

Rank	Native keywords	Frequency
1	Photovoltaic	301
2	Photovoltaics	154
3	Solar cells	119
4	Solar energy	115
5	MPPT	99
6	Renewable energy	89
7	Maximum power point tracking	69
8	Photovoltaic system	68
9	Solar cell	67
10	photovoltaic	62
11	photovoltaics	54
12	Photovoltaic systems	52
13	Silicon	48
14	Inverter	47
15	Efficiency	38
16	Photovoltaic cells	38
17	Photovoltaic (PV)	37
18	solar cells	36
19	solar cell	32
20	Organic solar cells	31
(…)	(…)	(…)
313	Kesterite	5
(…)	(…)	(…)
3971	kesterites	1
(…)	(…)	(…)
4375	kesterite	1
(…)	(…)	(…)
9737	Kesterites	1
(…)	(…)	(…)

Table 4
Top most frequently occurring clean keywords.

Rank	Clean keyword	Frequency
1	photovoltaics	893
2	solar cells (SC)	367
(…)	(…)	(…)
9	thin film (TF)	130
10	photovoltaic cell	121
(…)	(…)	(…)
13	dye sensitized solar cell (DSSC)	91
(…)	(…)	(…)
16	organic solar cells (OSC)	70
(…)	(…)	(…)
21	quantum dots (QD)	57
(…)	(…)	(…)
27	photovoltaic thermal	49
(…)	(…)	(…)
31	fullerene	43
32	nanostructure	42
(…)	(…)	(…)
38	nanowire	39
(…)	(…)	(…)
42	Perovskite solar cell	38
(…)	(…)	(…)
140	intermediate band (IB)	16
(…)	(…)	(…)
199	pv modelling	12
(…)	(…)	(…)
332	kesterite	8
(…)	(…)	(…)

We consider the indicator described by (1):

$$r(N) = n(MK \cap N) \tag{1}$$

The value of r(N) is the number n of MKs present in the first N-ranked (most frequently-occurring) keywords retrieved by TIM under specified search/filtering conditions. This corresponds to the concept of “true positives” actually retrieved by the text-mining method, but at a “fixed ranking” (in the first N-ranked retrieved results) since the cognitive selection is performed by the analysts only for the first N elements retrieved by the software (2.2.4). The specific r(300) indicator (for the first 300 most frequently-occurring keywords) is considered more suitable for the purposes of this paper, while r(100) (for the first 100 keywords) is deemed to retrieve insufficient MKs (see section 4).

The indicator in (2):

$$Recallrate(N) = \left(\frac{r(N)}{MK} * 100\right) \tag{2}$$

is the percentage of MKs present among the first N-ranked keywords retrieved by TIM under specified search/filtering conditions. This could be also defined “Recall rate at a fixed ranking”. We calculated Recallrate(300) and Recallrate(100) for different TIM settings (examples in section 4).

We identify as Rank(MK_i) the value of the rank of a specific marker keyword in a list of keywords produced by the software, under specified search/filtering conditions. For example, the rank of the marker keyword “Organic solar cells” (third row in Table 1) in the list “native keywords” (Table 3) is 20.

On the base of this it is possible to define the function:

$$SumRank(MK) = \sum_{i=1}^{n(MK)} Rank(MK_i) \tag{3}$$

This function is the sum of the ranks of all the MKs retrieved by the software under specific search/filtering conditions. This indicator quantifies the success of the software in high-ranking a set of marker keywords: the lower this indicator the higher the efficacy. This indicator can be calculated only if all the keywords of the MK set are present also in the list produced by the software.

Considering a possible failure of the software in identifying all the MKs (or a “loss of information”, compared to the expert's review) it is necessary to define a similar indicator only for a subset (MK-W) of marker keywords, by removing the “W” worst-performing marker keywords in the considered keyword list (those with highest values of the rank, or not present in the list), as in equation (4):

$$SumRank(x\%) = SumRank(MK - W) = \sum_{i=1}^{n(MK-W)} Rank(MK_i) \tag{4}$$

The indicator in (4), compared to that in (3), can allow rank sum calculations even if a maximum number of “W” MKs are missing from the list under exam. The number of the W worst performing MKs which could be excluded from the calculation should be an amount considered a “reasonable” loss of information. For the purposes of this paper, we consider as “satisfactory” the ability to bibliometrically retrieve about 68% of the MKs designated by an expert review (the percentage of values within two standard deviations of the mean in a normal distribution). The “reasonable loss of information” would then be about 33% in our population of 24 MK, or 8 marker keywords. Conversely, since the SumRank function is calculated on the top 66% of retrieved MKs, we can call it SumRank (66%).

Section 4 presents some numerical application examples.

3. Results: keyword lists generated by TIM

As outlined in 2.2.3, TIM uses a Boolean search string to retrieve scientific documents (papers, projects, patents) from which it extracts and cleans keywords relevant to the concept of emerging PV technologies. below list the most frequently occurring native and clean keywords.

3.1. Native list of keywords

Table 3 lists the top 20 most frequently occurring native keywords, for the years 1996–2016, with additionally the example of kesterite. TIM recovered 9770 native keywords, which appear as they were input by authors/editors, rank-ordered from most to less frequently occurring. Variants such as “Solar cells” and “solar cell”, or “Kesterite” and “kesterites” appear and are ranked as distinct keywords. The sum of the number of keyword occurrences (cumulated frequency) is about 16 000.

3.2. Post clumping or cleaned list of keywords

As outlined in section 2.2.3, TIM applies a term clumping process to the native keywords listed in Table 3, thereby harmonising and grouping words with similar morphological roots (e.g. same name in uppercase, lowercase etc.) into a single concept (e.g. “solar cell” or ”kesterite”). TIM uses a dictionary, with synonyms and acronyms from scientific literature. After this reduction step 5795 clean keywords are left (Table 4 presents a selection), down from the 9770 of Table 3.

The cumulated frequency is mainly the same before and after the cleaning process, because TIM groups keywords according to the processing described in 2.2.3 and adds up their frequency of occurrence. For example, in the native keyword list we have: “Kesterite”: 5 occurrences (rank 313 of Table 3); “kesterites”: 1 occurrence (rank 3971); “kesterite”: 1 occurrence (rank 4375) and “Kesterites”: 1 occurrence (rank 9737). These are grouped into the single clean keyword “kesterite”, with 8 occurrences (rank 332 in Table 4). As the dictionary is not based on a PV-specific corpus of documents, some terms denoting the same domain-specific concept (e.g. “solar cells” (rank 2 in Table 4) and “photovoltaic cell” (rank 10) still appear as distinct keywords in the clean keyword list.

3.2.1. Keyword rank-frequency plots

A rank-frequency chart (Fig. 5) was prepared based on clean keywords exported from TIM for three time periods: 1996 to 2005 (triangular ▲ marks), 2006 to 2010 (circular ● marks) and 2010 to 2015 (diamond ◆ marks)10TR8GXPF01. The data plotted in Fig. 5 are breakdowns of the data aggregated above for 1996 to 2015.

The three keyword rank-frequency plots are in log-log coordinates, with the horizontal axis showing the ranks of keywords in the frequency table, and the vertical axis indicating the total number (frequency) of the keyword's occurrences. The three plots approximately obey a heavy-tailed power law distribution and roughly follow commonly used (near-Zipf) models of term distributions, whereby frequency very rapidly decreases with rank [18].

The keyword rank-frequency distributions plotted in Fig. 5 have visibly different slopes in the head, middle and tail segments.

The steeper slope in the head segment can be largely attributed to the fact that similar highly-frequently occurring native keywords (e.g. Solar Cell, Solar Cells, solar cells) have been harmonised and grouped by TIM, as indicated in item 3.2, thereby further increasing

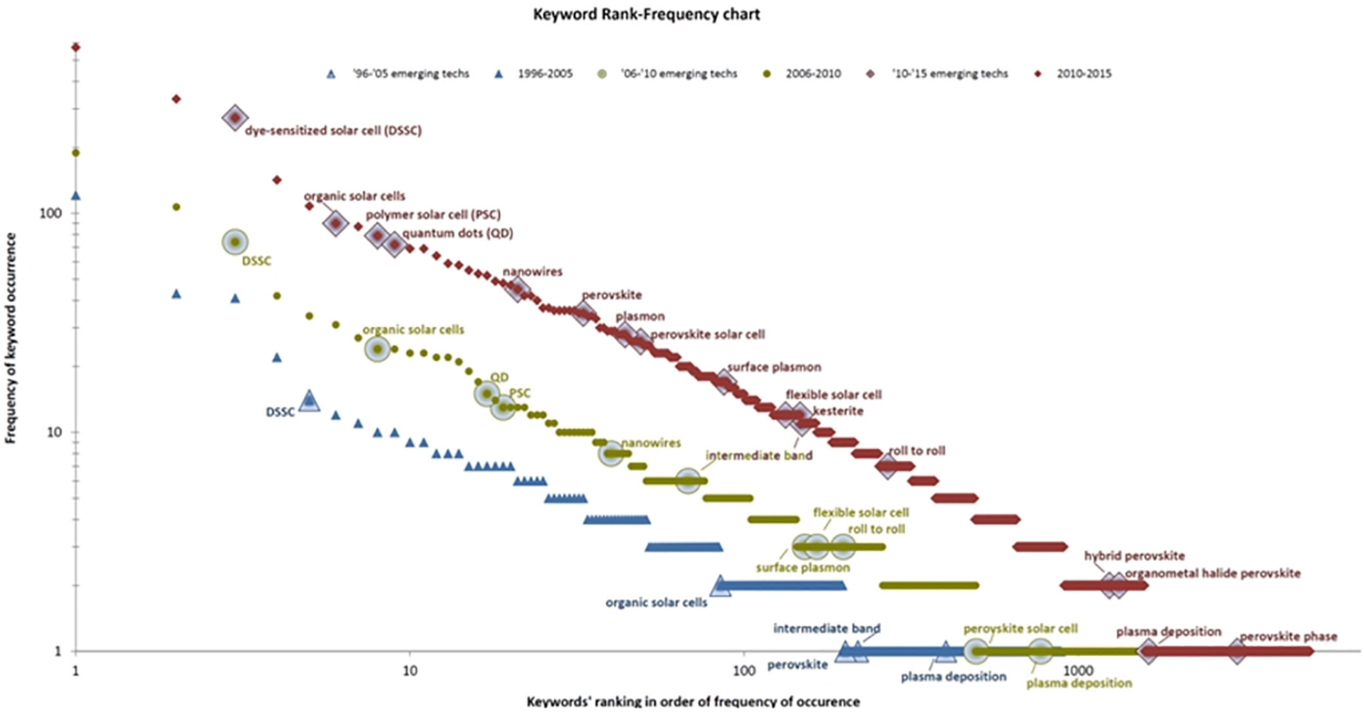

Fig. 5. Keyword rank-frequency chart for emerging PV technologies.

their frequency of occurrence. As shown above, TIM also retrieves keywords used in the search string (e.g. solar cell, photovoltaic), which are not considered to convey relevant information and are therefore omitted from the analysis. Other keywords appearing in the upper part of the distribution are too generic (e.g. thin film, renewable energy) and/or cross-sectoral (e.g. silicon) to be considered relevant for this analysis. The only keywords that appeared in both the head of the distribution and in the list of candidate emerging technologies identified by PV experts (Table 1) are ”dye sensitized solar cell” and “organic solar cells”, plus its equivalent “polymer solar cell”. These three keywords could be considered as the “master keywords” designating two technologies in the more “mature” part of the spectrum of emerging PV technologies, within the scope of this paper.

Regarding the middle part of the distribution, TIM retrieved nine of the technologies identified as emerging by the PV technology experts, and ranked them between 12 and 400 in terms of frequency of occurrence in the entire period from 1996 to 2015 (Fig. 5). This roughly confirms the analysts' intuitive approach of manually checking the first most frequently occurring 300 keywords as mentioned above in section 2.2.4. It is also in line with the established understanding that mid-range terms are the best index terms and relevance discriminators [30,31].

4. Benchmarking keyword lists generated by TIM against experts

Section 3 presented results from the TIM-based bibliometric analysis. For the purposes of this paper, it is important to test/benchmark these results against findings from expert reviews. In this section, we do so by assessing the effectiveness of TIM-based bibliometrics in retrieving “marker keywords” (MKs) designated by international experts (section 2.1) as representing emerging technologies in the photovoltaic sector.

Section 2.3 described the indicators we intend to use to assess the keyword retrieval effectiveness of TIM-based bibliometrics. The TIM software was run several times using the same search string described in section 2.2 but under different filtering/setting conditions.

Automated keyword-cleaning effectiveness by TIM was quantitatively assessed (Table 5) by comparing results for:

- Native + all + 20y (reference case): TIM native keyword list (section 3.1); all document sources (SCOPUS, CORDIS, PATSTAT); 20 years (documents issued from 1996 to 2016)
- Clean + all + 20y: TIM-filtered “cleaned keyword list” (section 3.2); all document sources; 20 years.

Table 5
Automated keyword-cleaning efficacy of TIM.

Indicators	TIM settings	
	Native + all + 20y	Clean + all + 20y
Number of keywords retrieved by TIM	9770	5795
r(300) or MKs retrieved in the first 300	12	15
Recallrate(300)	50%	63%
SumRank(MK)	29671	N.a.
SumRank(66%)	3113	1982
r(100) or MKs retrieved in the first 100	9	8
Recallrate(100)	38%	33%

Note: MK = Marker Keywords (24) identified by experts.

The effectiveness of varying the timespan of the bibliometric searches was assessed for our purposes by comparing results (Fig. 6) for the settings:

- Reference case (see above)
- Native + all + 10y: native keyword list; all document sources: 10 years (2006–2016).
- Native + all + 8y: native keyword list; all document sources: 8 years (2008–2016).
- Native + all + 6y: native keyword list; all document sources: 6 years (2010–2016).
- Native + all + 4y: native keyword list; all document sources: 4 years (2012–2016).

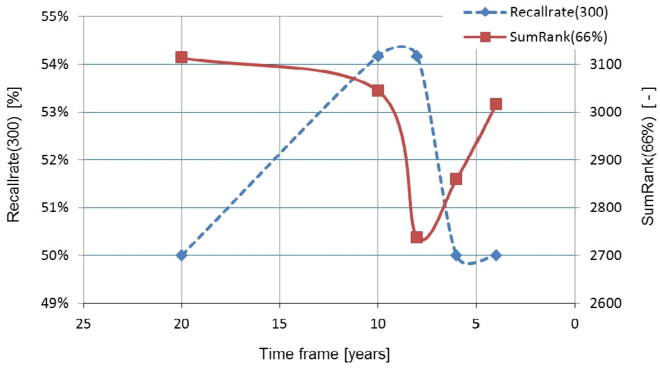

Fig. 6. Performances of the TIM software by varying the time frame.

We also looked at the effect of running the searches for scientific publications only (Table 6):

- Reference case (see above)
- Native + papers + 20y: native keyword list; SCOPUS, as document source, considering only papers from journals and conference proceedings; 20-year time frame.

Table 5 shows quantitative values of the indicators defined in section 2.3, calculated for the specified settings. In the native-keyword list, among the 300 most frequently occurring (highest-ranked) keywords, 50% of the marker keywords from the experts are retrieved by TIM (Recallrate(300) = 50%). The TIM automated cleaning algorithm (section 2.2.3) further improves this performance: the third column of Table 5 shows the Recallrate(300) rising to 63%.

Also the parameter SumRank(66%), summing the ranks of the top 66% MKs retrieved by TIM, improves for the “clean keyword” filter (SumRank indicators best perform when lower: see section 2.3). Calculations for the indicators r(100) and Recallrate(100) (information retrieval among the first 100 most frequently occurring keywords) indicate (Table 5) that 100 is an insufficient number of keywords to retrieve a meaningful amount of relevant information from bibliometric searches as we performed them.

Fig. 6 shows results for varying the time frame of the searches. The values of Recallrate(300) indicate that filtering the search results by reducing the time frame to the 10 and 8 most recent years increases the recovery of marker keywords, compared to the reference setting of 20 years. This is also confirmed by the improvement of the indicator Sum_Rank(66%), which performs best (the lower the better) for a time frame of 8 years. We attribute this improvement to the fact that technologies have a more consolidated jargon in recent than in previous years, so retrieved keywords have higher frequencies of occurrence. Alternatively, overly-reducing the time frame leads to loss of information as shown by lower recall rate values for 6 and 4 years. The best performing time frame for the searches discussed in this paper is 8–10 years. It is possible to notice that analysing the TIM results by mean of the indicator Sumrank(66%) provides more information (five different values: 3113, 3044, 2738, 2859, 3016 with a good variability) than the indicator Recallrate(300), providing only two values (50% and 54%).

Regarding document sources, the main information is seen to come from journals and conference proceedings found in the SCOPUS database, since Recallrate and Sumrank(300) are performing equally (Table 6).

The output of these tests that we performed on the TIM software can be considered very good, because TIM retrieved 63% of the technologies identified by experts at a cost of less than 10% of the cost of an expert elicitation and in a time that is similarly in the order of 1/10 of the time necessary to organise the expert elicitation event.

Table 6
TIM performances for all the document sources Vs papers from SCOPUS.

Indicators	TIM settings	
	Native + all + 20y	Native + papers + 20y
Number of keywords retrieved by TIM	9700	2949
r(300) or MKs retrieved in the first 300	12	12
Recallrate(300)	50%	50%
SumRank(MK)	29671	N.a.
SumRank(66%)	3113	3124
r(100) or MKs retrieved in the first 100	9	9
Recallrate(100)	38%	38%

Note: MK = Marker Keywords (24) identified by experts.

5. Discussion

In horizon scanning, text-mining bibliometric methods can be cheaper and quicker than a classic fully-fledged expert review, typically involving tens of highly qualified experts. Running the expert review exercise used here as reference [5] to benchmark the TIM software keyword retrieval performance required not only budget (for external experts, facilities, etc.) but also human resources to plan and organise the review event. Moreover, the expert elicitation process can be quite long, especially if highly qualified – therefore very busy – experts are required. Once tools and procedures are available the bibliometric methods can be run in shorter time frames.

However, these methods are relatively new and, from our experience, it is fundamental to rely on experienced analysts and to have the support and feedback of experts in the field, as discussed above and highlighted in bibliography [13,8,10]. Kajikawa et al. [32] argue that “vision is usually given by a top-down approach based on experts' experience and intuition, but its feasibility should be tested against existing data and trends.”

5.1. Limitations of the use of bibliometric software to identify emerging technologies

The use of bibliometric software to retrieve keywords and identify emerging concepts/technologies has some intrinsic limitations that need to be considered, affecting not only TIM but bibliometric software relying on keywords in general. First, emerging concepts/technologies are intrinsically complex and cannot be identified with a single keyword; then, the authors/editors tend to fragment complex concepts into different keywords. This makes it difficult to identify, starting from a single keyword, whether the given keyword is part of a more complex concept or not. For example, technologies such as “nanoscale heterojunction” may be split into the two keywords “nanoscale” + “heterojunction”, which taken alone (and ranked differently) do not give a full picture. Therefore analysts trying to identify technologies starting from single keywords need to consider this and also pay attention to single keywords apparently out of context (“nanoscale”) but that can be implicitly linked to the name of a related technology.

This is much truer for emerging technologies, because upon emergence a technology often still does not yet have a specific recognised name and authors/editors tend to categorise it differently. This may reduce the frequency of occurrence of keywords, and therefore lower their ranking by the bibliometric software. For example, kesterite solar cells have been categorised for several years by their chemical composition, which can vary and can have different acronyms (Cu2ZnSn(S,Se)4, CZTS, CZTSe, CZTSS).

5.2. *Expert insights and bibliometric analysis*

The added-value of combining expert reviews and bibliometric analysis is well documented. For example, Rip and Courtial [27] refer to the need for analysts to "get a feel for the overall picture and the fine structure" of bibliometric maps, adding that "the computer programs facilitate experimenting with different indices and thresholds, but interpretation remains intuitive". They also argue that quantitative evidence from bibliometrics can "introduce some distance and quasi-objectified procedures for analysing scientific fields that can be checked by actors as well as analysts", thereby mitigating risks of cognitive bias and analyst-actor dilemmas sometimes associated with classic cognitive analyses.

Zhang et al. [12] note that "qualitative methodologies depend on experts" intuitive knowledge and they may be biased since the opinion of experts may be influenced by subjective elements and limited cognitive horizons. In the opposite case, quantitative methods indicate both actual and potential features from science, technology and innovation activity tabulation, document text mining and other data". As limitations of quantitative methods (e.g. bibliometrics), they mention that, for example: not all R&D is published or patented and counts do not distinguish quality; not all publications or patents are similarly valuable (e.g. patent barriers could have more business value than the technology itself; science, technology and innovation database coverage lags can be important in analysing search results.

5.3. *Technology ranking and Technology Readiness Levels*

The Technology Readiness Level (TRL) is an indicator useful to assess the degree of development of technologies [6]. One of the outputs required to the international expert panel was the estimation of a TRL for each technology, so as to allow analysing the relation between the frequency of occurrence (inversely proportional to the rank) of a MK retrieved by the bibliometric software and the TRL of the related technology. We observed that the frequency of occurrence of the retrieved keywords (or, in relative terms, the order of magnitude of their rank in the list) often corresponds to the degree of development of the related technology: more mature technologies are ranked higher than emerging ones. In the clean keyword list (Table 4) the most frequently occurring MKs (with the best ranks) are "dye-sensitised solar cells" (# 13) and "organic solar cells" (#16), which according to the experts have the highest TRL (5–6). On the other hand, most keywords designating technologies with lowest TRL (1–2) appeared less often and therefore were ranked in the lowest positions. For example, "innovative III-V compound solar cells" was ranked 211, "photocatalysis" ranked 226, "ferroelectric PV" was ranked 285. "perovskite" also conforms to this trend, having average TRL (4–5) and average rank (#42). This behaviour confirms what was said in 5.1. Even if we searched for "emerging" technologies, at higher TRLs technologies tend to have more consolidated jargon. Keywords designating more mature technologies tend to appear more often and therefore be more retrievable by bibliometric frequency analysis methods.

5.4. *Market perspectives for emerging PV technologies*

Wafer-based silicon solar cells had, in 2016, a market share of 90% and continue to be the main PV technology. Thin film, the main competitor of wafer silicon, reached a market share of 20% in 2009 and then decreased. This is related to a limited progress in the developmet of these technologies and a lack of investments in start-up companies. On the research level, new non-concentrating high efficiency concepts are increasing the cell and module efficiencies (Table 1). The existing PV technology mix is the base for the future growth of the sector as a whole. No single technology can satisfy all the different consumer requirements, ranging from mobile and consumer applications, and the need for a few watts up to multi-MW utility-scale power plants. If material limitations or technical obstacles restrict the further growth or development of a technology pathway, then the variety of technologies will be an insurance against possible barriers in the exploitation of solar PV electricity [33]. The current perspective for the emerging technologies is to find a niche market, not to compete with the well estabilshed silicon market. TRLs for the most advanced emerging technologies (DSSC, organic) seem to be progressing quite slowly. However, research on these fields has paved the way for perovskite solar cells, which experienced an impressive efficiency evolution in the last ten years and are now attracting attention of the scientific community and the industry [5].

6. Conclusions

Bibliometric software can be used to help policymakers identify emerging and promising technologies in specific sectors, by means of quantitative analysis algorithms.

This paper demonstrates the use of Tools for Innovation Monitoring (TIM), a bibliometric analysis software recently developed by the JRC for this purpose, including a case study with quantitative examples of its use to identify emerging technologies in the field of photovoltaics.

The outputs of this software (keyword lists) are compared with qualitative cognitive analysis by an expert review by means of specifically tailored indicators.

TIM was tested under different setting/filtering conditions, showing satisfactory results. Its automated filtering (clumping) function was effective: about 63% of the 24 technologies identified by experts as relevant emerging PV technologies were ranked by TIM among the first 300 keywords it retrieved. A time frame of 8 years was found the most relevant for searching emerging technologies in the frame of this setup. Scientific articles and proceedings from the SCOPUS database were prominent among the search results.

This bibliometric software retrieved most of the technologies identified by experts at a cost of 10% of the cost of the expert elicitation, and in a time that is similarly in the order of 1/10 of the time necessary to organise the expert elicitation event.

The most salient technologies (identified amid the most frequently occurring keywords) tend to be more consolidated technologies, with higher Technology Readiness Levels. Lower TRL technologies (TRL between 1 and 3) tend to occur less frequently and therefore be more difficult to be retrieved in the higher-rank positions.

The JRC is further developing the TIM Tools for Innovation Monitoring and expanding the analysis to other emerging energy technologies.[1]

Acknowledgements

The authors gratefully acknowledge the contribution of Sergio Perani and all the team working on the development of the TIM software, Arnulf Jaeger-Waldau for his support as in-house senior

[1] http://timanalytics.eu/.

expert on identifying emerging PV technologies, Nigel Taylor for his comments, insights and encouragement, David Shaw for his erudite review and three unknown reviewers for their comprehensive work.

References

[1] European Commission, Communication from the commission to the European parliament, the council, the European economic and social committee and the committee of the regions, Build. Eur. Data Econ. (2017). https://ec.europa.eu/taxation_customs/sites/taxation/files/1_en_act_part1_v10_en.pdf.

[2] G. Joanny, A. Agocs, S. Fragkiskos, N. Kasfikis, J-M Le Goff, Monitoring of Technological Development - Detection of Events in Technology Landscapes through Scientometric Network Analysis, in: Proc. ISSI 2015 Istanbul 15th Int. Soc. Sci. Inf. Conf. Istanbul, Turkey, 29 June to 3 July, 2015 1130–1141, Boğaziçi University Printhouse, 2015.

[3] E. Boelman, T. Telsnig, R. Shortall, G. Bardizza, A. Villalba Pradas, Bibliometric network densification patterns for three renewable energy technologies, in: GTM2017 7th Glob. TechMining Conf, 2017.

[4] E. Amanatidou, M. Butter, V. Carabias, T. Könnölä, M. Leis, O. Saritas, P. Schaper-Rinkel, V. van Rij, On concepts and methods in horizon scanning: lessons from initiating policy dialogues on emerging issues, Sci. Public Policy (2012), https://doi.org/10.1093/scipol/scs017.

[5] A. Moro, J. Aycart, G. Bardizza, M. Bielewsky, J. Lopez-Garcia, N. Taylor, A.F. Lazo, L.J. Garcia, First workshop on identification of Future Emerging Technologies for low carbon Energy Supply, 2017, https://doi.org/10.2760/849373.

[6] European Commission, Technology readiness levels (TRL), 2015. https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-g-trl_en.pdf.

[7] L. Huang, Y. Zhang, Y. Guo, D. Zhu, A.L. Porter, Four dimensional Science and Technology planning: a new approach based on bibliometrics and technology roadmapping, Technol. Forecast. Soc. Chang. (2012), https://doi.org/10.1016/j.techfore.2012.09.010.

[8] F. Rizzi, N.J. van Eck, M. Frey, The production of scientific knowledge on renewable energies: worldwide trends, dynamics and challenges and implications for management, Renew. Energy 62 (2014) 657–671, https://doi.org/10.1016/j.renene.2013.08.030.

[9] L. Leydesdorff, Can technology life-cycles be indicated by diversity in patent classifications? The crucial role of variety, Scientometrics (2015), https://doi.org/10.1007/s11192-015-1639-x.

[10] J. Zhang, Y. Yan, J. Guan, Scientific relatedness in solar energy: a comparative study between the USA and China, Scientometrics (2014), https://doi.org/10.1007/s11192-014-1487-0.

[11] N.J. Van Eck, L. Waltman, VOSviewer: A Computer Program for Bibliometric Mapping, ERIM Report Series Reference No. ERS-2009-005-LIS, November 2, 2009, p. 2009. Available at SSRN: https://ssrn.com/abstract=1346848.

[12] Y. Zhang, Y. Guo, X. Wang, D. Zhu, A.L. Porter, A hybrid visualisation model for technology roadmapping: bibliometrics, qualitative methodology and empirical study, Technol. Anal. Strateg. Manag. (2013), https://doi.org/10.1080/09537325.2013.803064.

[13] A.L. Porter, Y. Zhang, Tech mining of science & technology information resources for future-oriented technology analyses, in: J.C., G. Glenn (Eds.), Futur. Res. Methodol. Version 3.1, the Millennium Project, 2015. http//themp.org/.

[14] L. Georghiou, J.C. Harper, M. Keenan, I. Miles, R. Popper, The Handbook of Technology Foresight : Concepts and Practice, 2008.

[15] Elsevier, Scopus, Scopus [Database], 2017.

[16] European Commission, CORDIS: Community Research and Development Information Source, 2017 (Accessed 27 January 2018), http://cordis.europa.eu/.

[17] European Patent Office, PATSTAT: Worldwide Patent Statistical Database, 2017 (Accessed 27 January 2018), https://www.epo.org/searching-for-patents/business/patstat.html#tab-1.

[18] C.D. Manning, P. Raghavan, H. Schutze, An introduction to Information Retrieval, 2009, https://doi.org/10.1109/LPT.2009.2020494.

[19] G. Vidican, W. Woon, S. Madnick, Measuring innovation using bibliometric techniques: the case of solar photovoltaic industry, Sloan Work. Pap (2009), https://doi.org/10.2139/ssrn.1388222.

[20] G. Prathap, A three-dimensional bibliometric evaluation of research in polymer solar cells, Scientometrics (2014), https://doi.org/10.1007/s11192-014-1346-z.

[21] A.L. Porter, J. Youtie, P. Shapira, D.J. Schoeneck, Refining search terms for nanotechnology, J. Nanoparticle Res. (2008), https://doi.org/10.1007/s11051-007-9266-y.

[22] 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 (2013), https://doi.org/10.1007/s11192-012-0903-6.

[23] M. Sahlgren, R. Cöster, Using bag-of-concepts to improve the performance of support vector machines in text categorization, in: Proc. 20th Int. Conf. Comput. Linguist. - COLING '04, 2004, https://doi.org/10.3115/1220355.1220425.

[24] O. Täckström, An evaluation of bag-of-concepts representations in automatic text classification, Recall (2005).

[25] H.K. Kim, H. Kim, S. Cho, Bag-of-concepts: comprehending document representation through clustering words in distributed representation, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.05.046.

[26] P.C. Lee, H.N. Su, Investigating the structure of regional innovation system research through keyword co-occurrence and social network analysis, Innov. Manag. Policy Pract. (2010), https://doi.org/10.5172/impp.12.1.26.

[27] A. Rip, J.P. Courtial, Co-word maps of biotechnology: an example of cognitive scientometrics, Scientometrics (1984), https://doi.org/10.1007/BF02025827.

[28] W. Li, Y. Zhao, Bibliometric analysis of global environmental assessment research in a 20-year period, Environ. Impact Assess. Rev. (2015), https://doi.org/10.1016/j.eiar.2014.09.012.

[29] M. Sanderson, Test collection based evaluation of information retrieval systems, Found. Trends® Inf. Retr. (2010), https://doi.org/10.1561/1500000009.

[30] R.M. Losee, Term dependence: a basis for Luhn and Zipf models, J. Am. Soc. Inf. Sci. Technol. (2001), https://doi.org/10.1002/asi.1155.

[31] H.P. Luhn, The automatic creation of literature abstracts, IBM J. Res. Dev. 2 (1958), https://doi.org/10.1147/rd.22.0159.

[32] Y. Kajikawa, Y. Takeda, K. Matsushima, Computer-assisted roadmapping: a case study in energy research, Foresight (2010), https://doi.org/10.1108/14636681011035726.

[33] A. Jäger-Waldau, PV Status Report 2017, 2017, https://doi.org/10.2760/731933.