

Tracing technology evolution pathways by combining patent citation analysis and tech mining

Ying Huang^{1,2,4}, Yi Zhang^{1,2,3}, Jing Ma^{1,2}, Alan L. Porter^{4,5}, Xuefeng Wang^{*,1,2,4}

¹ School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China

² Lab of Knowledge Management and Data Analysis (KMDA), Beijing Institute of Technology, Beijing 100081, China

³ Decision Systems and e-Service Intelligence research Lab, Centre for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia

⁴ School of Public Policy, Georgia Institute of Technology, Atlanta, GA 30332, USA

⁵ Search Technology, Inc., Atlanta, GA 30092, USA

Abstract

Because of the flexibility and complexity of Newly Emerging Science and Technologies (NESTs), traditional statistical analysis fails to capture technology evolution in detail. Generating competitive technical intelligence supports industrial, governmental, and academic decisions to guide future development trends.

Patents are one of the most important NESTs data sources and are pertinent to developmental paths. This paper draws upon text analyses, augmented by expert knowledge, to identify key NESTs sub-domains and component technologies. We then complement those analyses with patent citation analyses to help track developmental progressions. We identify key sub-domain patents, associated with particular component technology trajectories, then extract pivotal patents via citation analyses. We compose evolutionary pathways by combining citation and topical intelligence obtained through term clumping.

We demonstrate our approach with empirical analysis of Dye-Sensitized Solar Cells (DSSCs). This study informs NESTs management by spotting prime opportunities for innovation.

Keywords: Innovation Pathways, Citation Analysis; Text Mining; Topic Analysis; Dye-Sensitized Solar Cells; Technology Roadmapping

I Introduction

Analyzing and tracking the historical and current stages of a technology are critical for gaining competitive advantage [4]. Patents, as a fruitful data source, are one main output of research and development that represent the characteristics of an emerging technology, and, thus, they are important for technology innovation management [11]. A large portion of recent technical

knowledge is available in patent documents, and the importance of exploiting this knowledge is constantly increasing. Additionally, the number of patents is rapidly increasing, and the pace of technological development is accelerating. Identifying core and emerging technologies is crucial for formulating technology strategies and policies that pursue promising technological opportunities [3].

Compared with traditional technologies, Newly Emerging Science and Technologies (NESTs) have tremendous innovation potential [23]. Domain experts may become less reliable due to increasing data and the fragmentation of technology domains [24]. Data-based methods offer an appealing alternative to expert opinion. However, some patent-analytic methods use only simple bibliometric indicators -- e.g., logistic growth curves [17]-- and compare the numbers of patents assigned to different entities -- e.g. nations, companies, inventors, citations and technological fields-- over time [2,8,26]. Such indicators are useful, but they cannot reflect micro-level technology changes, especially for NESTs that manifest an increasing technological complexity and a shortened technology lifecycle [9]. More researchers are making efforts to adopt advanced qualitative techniques, including morphological analysis [12,28,30], TRIZ [31,34], conjoint analysis [13,27,29] and technology roadmapping [5,10,14,15,33,35]. Experts' involvement enhances the effective identification of competitive technical intelligence. However, such expert-based methods depend highly upon the experts' knowledge, experience and motivation, while experts' biases and, sometimes, insufficient knowledge may create difficulties. If, however, domain experts can reflect upon knowledge derived from data mining, then more accurate conclusions can be obtained. Thus, our research combines expert-based methods with large-scale (text) data-based methods.

In this paper, we combine patent citation analysis and tech mining to generate competitive technical intelligence, which achieves a balance between data-driven and expert-influenced conclusions. "Tech mining" [21] is a multi-step process to analyze Science, Technology and Innovation (ST&I) information resources by using text mining, visualization, and communication tools [22]. Term clumping, a characteristic method of "Tech mining", cleans and consolidates topical content in the publication and patent records and then extracts topical content intelligence [34]. Meanwhile, we explore patent development paths in a large-scale patent citation network by evaluating the weight of citations among patents to provide a more robust understanding of influential nodes. The combination of these methods can, to some extent, mitigate their respective drawbacks and make full use of their strengths in: 1) obtaining technical core terms in domain areas; 2) identifying influential nodes of a directed citation network; and 3) discovering significant clues about technology hotspots now and technology development prospects in the future.

Our case study focuses on Dye-Sensitized Solar Cells (DSSCs), a third-generation photovoltaic technology. The outstanding features of this technology-- e.g. low cost, ease of fabrication, environmental friendliness of raw materials, and relatively high efficiency -- have attracted tremendous scientific and industrial attention.

The remainder of this paper consists of four sections. In the section entitled "Data and Methodology," we describe the dataset and framework of our research. The next section, "Empirical Research: Competitive Technical Intelligence for DSSCs" applies the proposed approach for developing technology evolution pathways, and this incorporates patent topical analysis, patent citation analysis and term clumping analysis. Finally, in "Conclusions," we present remarks and directions for further study.

II Data and Methodology

In this research, we offer a systematic approach to trace technology evolution. The framework, organized in five steps, is illustrated in Figure 1. The five steps are as follows: (1) Download related patents (data search and collection); (2) Acquire sub-technology datasets; (3) Obtain core technological keywords; (4) Obtain complementary terms; and (5) Trace technology evolution roadmap.

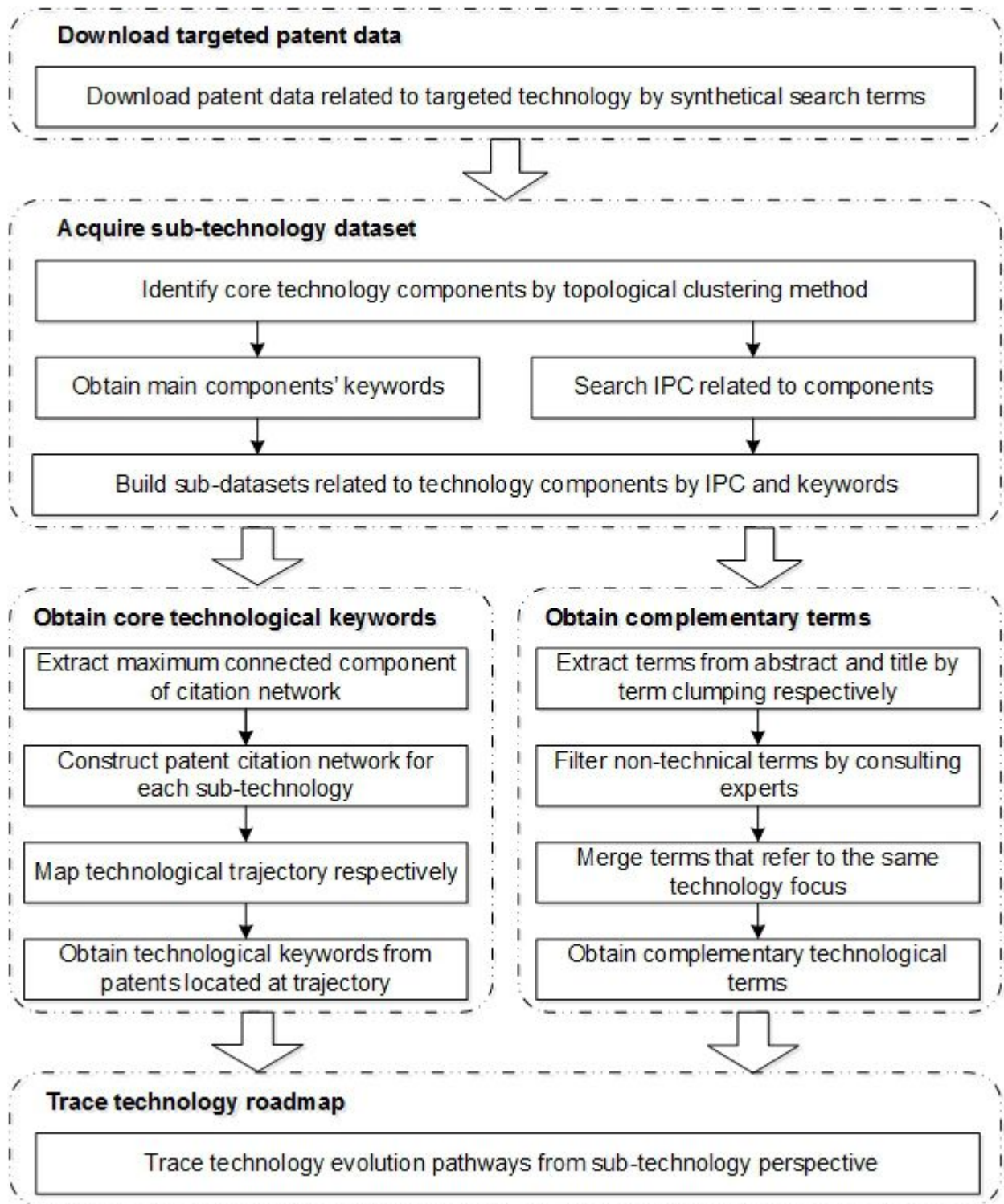


Fig. 1 Framework for tracing technology evolution

Data search and data collection

As the initial step of bibliometric analysis, devising an appropriate search strategy requires attention to assure quality data. However, it is notoriously difficult to define the boundary of a multidisciplinary and emerging field, and to harvest the relevant publications and patents. In our opinion, the search strategy should balance between precision and recall. On one hand, we need to retrieve a dataset whose records are relevant to the targeted field. On the other hand, we cannot spend too much time refining the search strategy to eliminate noise: after a certain point, the retrieval results become asymptotically stable, despite additional efforts to refine the search. We also recognize the potential to clean data, as needed, after download; so recall takes some priority over precision in our approach.

In this paper, we choose DSSCs as our target technology field. All patents are collected from the Thomson Reuters Derwent Innovation Index (DII) database. The search terms used are shown in set "#1" in Table 1. Unlike previous research, we combine keywords with manual codes in order to achieve more effective results. Keywords are easy to understand, even for those who are not specialized in the field. Manual codes, applied by a subject expert, not only cover standard technology terms, but they are also able to represent a technology and reflect the group (e.g. a patent family, a group of related inventions filed in one or multiple patent authorities) that each record belongs to [25]. Our search resulted in 5668 records, reflecting a group of related inventions filed in one or more patent authorities from 1991 to 2014 (retrieved on March 12, 2015).

Table 1. Search result of different sub-technologies of DSSCs

<i>Set</i>	<i>Number</i>	<i>Result</i>
# 1	5668	TS=(((dye-sensiti*) or (dye* same sensiti*) or (pigment-sensiti*) or (pigment same sensiti*) or (dye* adj sense)) same ((solar or Photovoltaic or photoelectr* or (photo-electr*)) same (cell or cells or batter* or pool*))) AND MAN=(L03-E05B1 OR U12-A02A8 OR X15-A02D1 OR X16-A04)
# 2	>100,000	IP=(H01L*) AND TS=(dioxide OR oxide) AND (anode))
# 3	2181	#2 AND #1
# 4	31960	IP=(C09B*) OR TS=(sensitiser OR sensitizer)
# 5	615	#4 AND #1
# 6	74849	IP=(H01G* OR H01M*) and TS=(electrolyte)
# 7	1674	#6 AND #1
# 8	97287	IP=(H01L-031/0224) OR (IP=(H01M*) AND TS=(electrode))
# 9	2348	#8 AND #1
# 10	53	#9 AND #7 AND #5 AND #3
# 11	4089	#9 OR #7 OR #5 OR #3

Note: Indexes=CDerwent, EDerwent, MDerwent Timespan=1991-2014

Topical analysis for sub-technologies

Unlike our previous work, in which we treated the domain technology as a whole, this paper divides the domain technology into sub-components for further exploration and more detailed

technology intelligence. International Patent Classification (IPC) terms provide a hierarchical system of language-independent symbols for classifying patents according to the different areas of technology to which they pertain (<http://www.wipo.int/classifications/ipc/>). When seen as technology classes, IPCs are useful analytical units for exploiting the information in patent databases [6].

After downloading the patents, we use *VantagePoint*, a professional text mining software (www.theVantagePoint.com), for topical clustering analysis. Natural Language Processing (NLP) techniques help extract a set of phrases and terms from specified textual fields (e.g. title and abstract). Next, we use ClusterSuite, a compilation of VantagePoint algorithms, to reduce noise, consolidate related items, and provide more refined topical information [20]. We apply Principal Components Analysis (PCA) to the top 200 informative phrases to identify the core technology components that are most often considered sub-technologies. One of the most important aspects of this research is record selection. In DSSCs, some high-frequency terms (e.g. semiconductor, solar cell) and common terms (e.g., photoelectric conversion, optoelectric transducer) should be removed. During the keyword selection process, we achieve better results with the help of domain experts. Thus, in this paper, related keywords and IPCs correspond to technology components that are identified with the help of domain experts, allowing us to obtain sub-technology datasets.

Patent citation analysis

Based on the sub-technology datasets, we construct the patent citation network consisting of both connected components and isolated nodes, respectively. We ignore the patents that never cite others or are never cited by others, and concentrate on the maximally connected component. We conduct this analysis in the following four steps:

- (1) Merge patents into record families. A patent family is the collection of patents in different countries referring to the same technical topic [18]. Our first step is to merge the patent documents of a family into a single family record, which is identified by the priority patent number (the earliest patent in this family).
- (2) Construct the patent citation network. While conducting main path analysis for a given field of technology, only citations between patents within the technology field are considered. These effective citations are extracted from the merged family records. A patent citation network can be represented as a patent citation matrix. Nodes stand for the individual family records, and the arcs between two nodes are citations.
- (3) Identify the main paths of the patent citation network. In a citation network, the main trajectory is the path from a source vertex to a sink vertex that has the highest traversal weight on its arcs. Several methods have been proposed to extract main paths from the network of traversal weights. The method we use is "Search Path Count" (SPC) [1], which extracts the main trajectory, meaning the path from a source vertex to a sink vertex with the highest traversal weight on its arcs.
- (4) Extract key technical intelligence from the patents located on the main path. The key patents located on the trajectory are obtained and their technological keywords are extracted manually. This information is used to create an initial technology evolution pathways.

Term clumping analysis

Although we have obtained the preliminary technology evolution pathways with patent citation analysis, the technology focus included in some less-cited patents might influence the whole technology development circle. One solution depends on the "terms" derived from NLP techniques. Phrases and terms retrieved in this way are large and "noisy," making them difficult to manually categorize. Using bibliometric and text mining techniques, this paper applies the semi-automatic "Term Clumping" steps, which generate better term lists for achieving competitive technical intelligence [33].

Term clumping includes four main phases: (1) Common and basic term removal, e.g., instance, technology; (2) Fuzzy word matching (combine terms with similar structures based on pattern commonality, such as stemming -- e.g., sensitiser and sensitizer, and combine singular and plural forms of English words, e.g., dye and dyes); (3) Extreme words removal [remove very common (top 5%) and very rare (occurrence in single records)]; (4) Combine term networks (combine select low-frequency phrases with the high-frequency phrases that appear in the same records, sharing terms).

After extracting the terms and phrases from titles and abstracts, some weakly correlated terms are removed after consulting with experts. At the same time, some keywords that indicate the same technology focus are merged to improve the integration level. The final keywords reflecting the technology focus are obtained to construct a technology evolution roadmap, building on former analysis experiences.

Empirical Research: Competitive Technical Intelligence for DSSCs

DSSC patent topical analysis

It is crucial to identify the main subsystems and the evolution roadmap for key topics of DSSCs. In the first step, we use VantagePoint's NLP to extract nouns and phrases to obtain a keyword list from the titles and abstracts of 5,668 records. 1,562 terms are obtained by the ClusterSuite process of term clumping analysis. In this context, we select the top 200 terms as the high level terms and use Factor Maps (PCA) to reduce the number of items for further topical analyses [19]. 17 clusters are generated, and most of them are related to each other. Based on experts' review, our preliminary cluster result effectively reflects the major characteristics of DSSCs. The four major sub-technologies are: Photoanode, Electrolyte, Counter-Electrode, and Sensitizer (Figure 2).

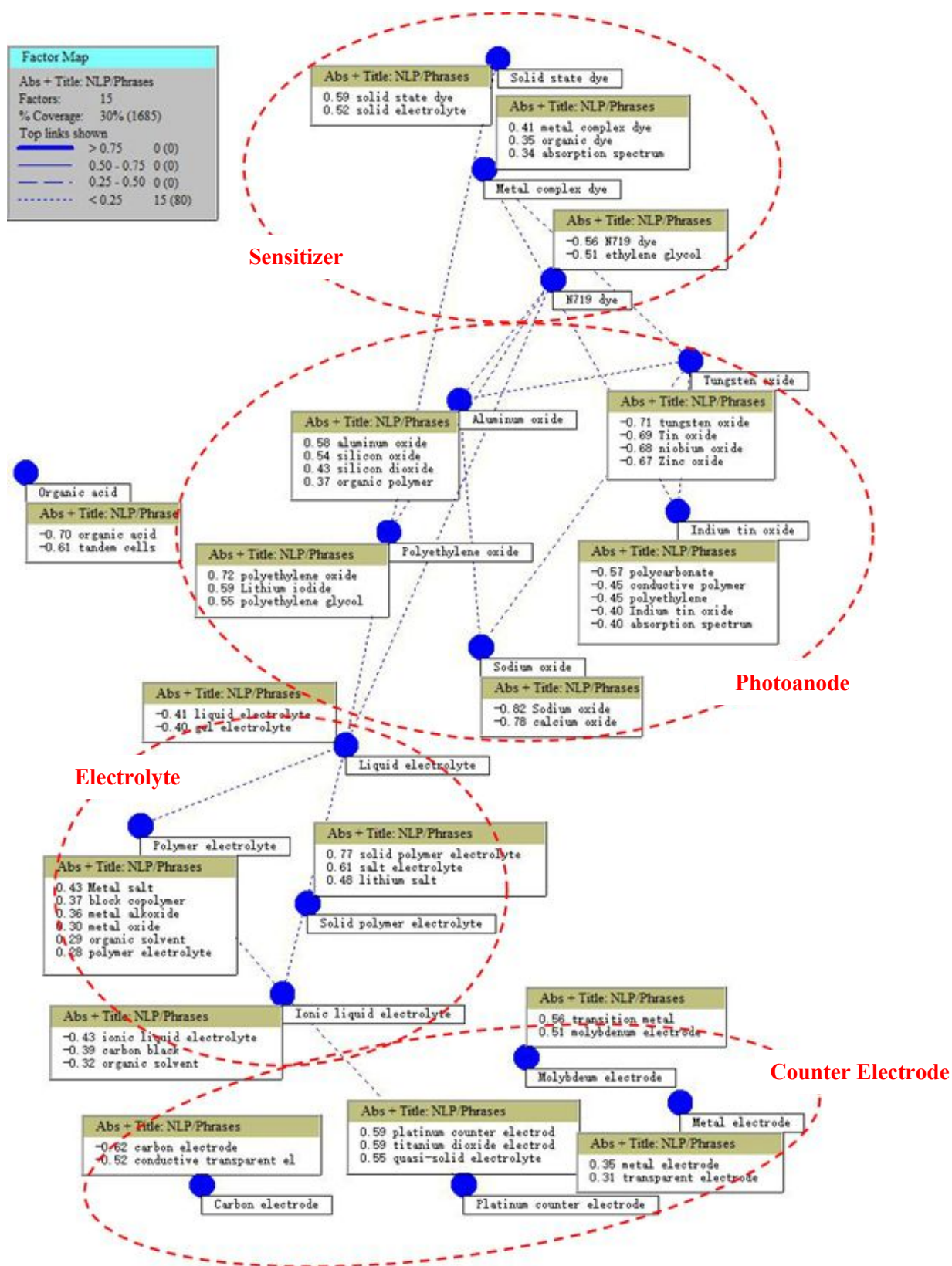


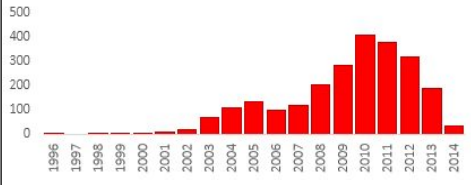
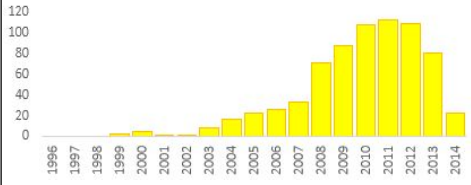
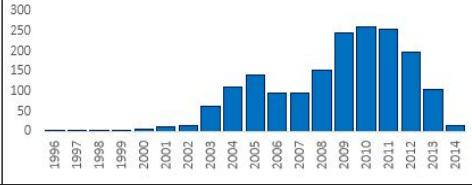
Fig. 2 Factor map of DSSCs (based on the top 200 topical terms).

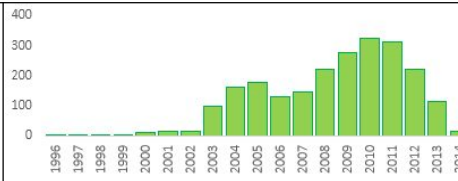
The photoanode components generally cover various nano-structured materials, such as titanium dioxide and zinc oxide, and these keywords often include "dioxide" or "oxide." Obviously, "anode" is another keyword. IPC H01L relates to semiconductor devices and semiconductor technologies, which are the key parts for photoanodes. In the sensitizers group, "dyes" is practically synonymous with "sensitizers," and C09B specializes in "organic dyes or closely-related compounds for producing dyes." For the electrolyte classification, H01G stands for electrolyte light-sensitive or temperature-sensitive devices, so most of the patents related to electrolyte are included. Additionally, because of the close relationship between electrolyte and counter-electrode, some patents are also located in the field of H01M, which mainly focuses on electrodes and voltage generators. Under the counter-electrode component, aside from H01M, there is a specialized classification named H01L-031/0224, which is related to counter-electrodes. The detailed search terms and their results are shown in Table 1.

We use keywords and IPCs to generate the sub-technology datasets from the original database. The advantage of this method is that it can get a more accurate dataset than the approaches that only use IPCs [32] or keywords [16]. The disadvantage of this method is the inherent risk that some patents are excluded. Overall, this approach helps us to obtain meaningful and useful, albeit incomplete, data in a reasonable way. In the end, 72.14% (4089 records) of the patents were divided into the four sub-technologies.

Table 2 represents the characteristics of the sub-technologies in DSSCs. Photoanode and counter-electrode are the two largest clusters among them, reaching 2181 and 2348 respectively, and both show similar growth curves. Sensitizer shows remarkably different growth curves when compared to the other sub-technologies; the patents drastically increase after 2008. In this paper, we calculate the average priority year of patents in each cluster to track emerging fields in DSSCs research. Among these four sub-technologies, photoanode and sensitizer are relatively new and emerging, and we speculate that more and more patents related to these two fields will appear.

Table 2. Characteristics of the sub-technology in DSSCs

#	Topics	Records	Keywords	Avg. year	Growth Curve
1	Photoanode	2181	Aluminum oxide; Tungsten oxide; Indium tin oxide; Polyethylene oxide	2009.21	
2	Sensitiser	615	Solid state dye; Metal complex dye; N719 dye	2009.83	
3	Electrolyte	1674	Liquid electrolyte; Polymer electrolyte; Ionic liquid electrolyte; Solid polymer electrolyte	2008.66	



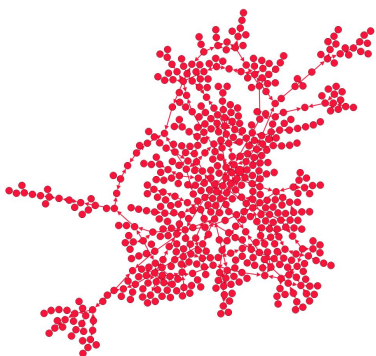
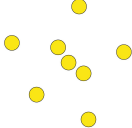
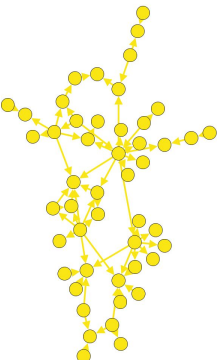
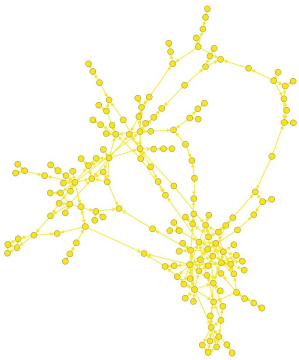

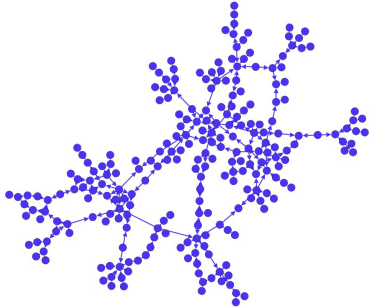
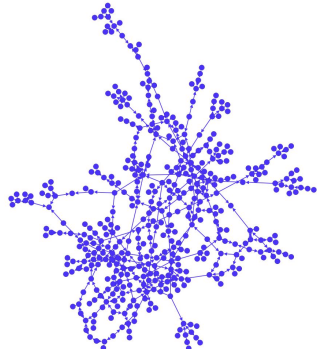

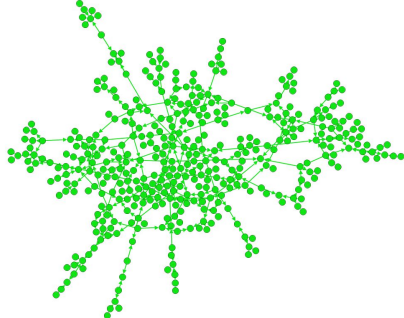
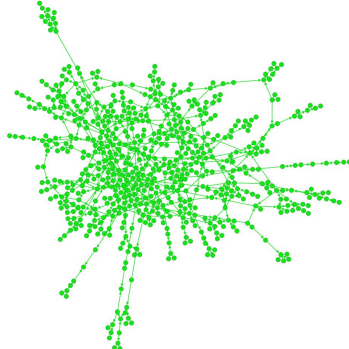
3	Counter-Electrode	2348	Carbon electrode; Molybdeum electrode; Platinum counter electrode; Metal electrode	2008.43	 <table><caption>Annual Patent Citations for Counter-Electrode (1996-2014)</caption><thead><tr><th>Year</th><th>Citations</th></tr></thead><tbody><tr><td>1996</td><td>0</td></tr><tr><td>1997</td><td>0</td></tr><tr><td>1998</td><td>0</td></tr><tr><td>1999</td><td>0</td></tr><tr><td>2000</td><td>0</td></tr><tr><td>2001</td><td>0</td></tr><tr><td>2002</td><td>0</td></tr><tr><td>2003</td><td>10</td></tr><tr><td>2004</td><td>150</td></tr><tr><td>2005</td><td>180</td></tr><tr><td>2006</td><td>120</td></tr><tr><td>2007</td><td>150</td></tr><tr><td>2008</td><td>220</td></tr><tr><td>2009</td><td>280</td></tr><tr><td>2010</td><td>320</td></tr><tr><td>2011</td><td>310</td></tr><tr><td>2012</td><td>220</td></tr><tr><td>2013</td><td>110</td></tr><tr><td>2014</td><td>10</td></tr></tbody></table>	Year	Citations	1996	0	1997	0	1998	0	1999	0	2000	0	2001	0	2002	0	2003	10	2004	150	2005	180	2006	120	2007	150	2008	220	2009	280	2010	320	2011	310	2012	220	2013	110	2014	10
Year	Citations																																												
1996	0																																												
1997	0																																												
1998	0																																												
1999	0																																												
2000	0																																												
2001	0																																												
2002	0																																												
2003	10																																												
2004	150																																												
2005	180																																												
2006	120																																												
2007	150																																												
2008	220																																												
2009	280																																												
2010	320																																												
2011	310																																												
2012	220																																												
2013	110																																												
2014	10																																												

DSSC patent citation analysis

The citation network of patents provides a representation of the innovation process [7]. In the development process, technology is present on different development tracks. Therefore, this context highlights the dynamic nature of technology in the development process in order to improve the accuracy of the analysis of sub-technologies in their technology evolution roadmaps. As previous research has shown, the whole of DSSC development can be divided into three stages: emerging stage (1991-2001), growth stage (2002-2009) and maturity stage (2010-2014)¹ [11]. In Table 3, it's not hard to see the evolution of DSSCs citation behaviors from the 1990s to 2014. We can see that the technology will continue to produce a more complex network of references as time goes on. This will include early inventions, which will be cited more frequently by later inventions. Likewise, recent patents will be cited by more patents in the future. In addition, analyzing the technology evolution will be a big challenge if only property analysis is used. Therefore, in this paper, we will extract the technology trajectory by means of the main path analysis.

¹ The previous research refine a maturity stage from 2010 to 2012 for choosing the time span from 1991 to 2012, but we choose the time span from 1991 to 2014 in this paper to obtain more newly data.

Table 3. Maximum patent citation networks of sub-technologies in DSSCs at different stages

	<i>1991-2001</i>	<i>1991-2009</i>	<i>1991-2014</i>
Photoanode			
Sensitizer			
Electrolyte			
Counter-Electrode			

The patents located on the main technological trajectory for sub-technologies in DSSCs from 1991 to 2014 are shown in Figure 3. For photoanode, there are 11 patents selected as the characteristic nodes to represent the evolution pathway, and its trajectory has three clusters. The sensitizer trajectory includes 8 patents and has an arrowhead-shaped citation path. For electrolyte, 13 patents are located on the technology trajectory where they produce a sudden change, driven by EP1865522, which takes advantage of the electrolyte characteristics therein. For counter-electrode, the technological trajectory includes 12 patents, among which is a Korean patent (Patent No. WO2007064164) that uses carbon nanotube electrode, which produces a much better result than the previous platinum electrode. From that point on, carbon counter-electrodes play an important role in improving photoelectric conversion efficiency.

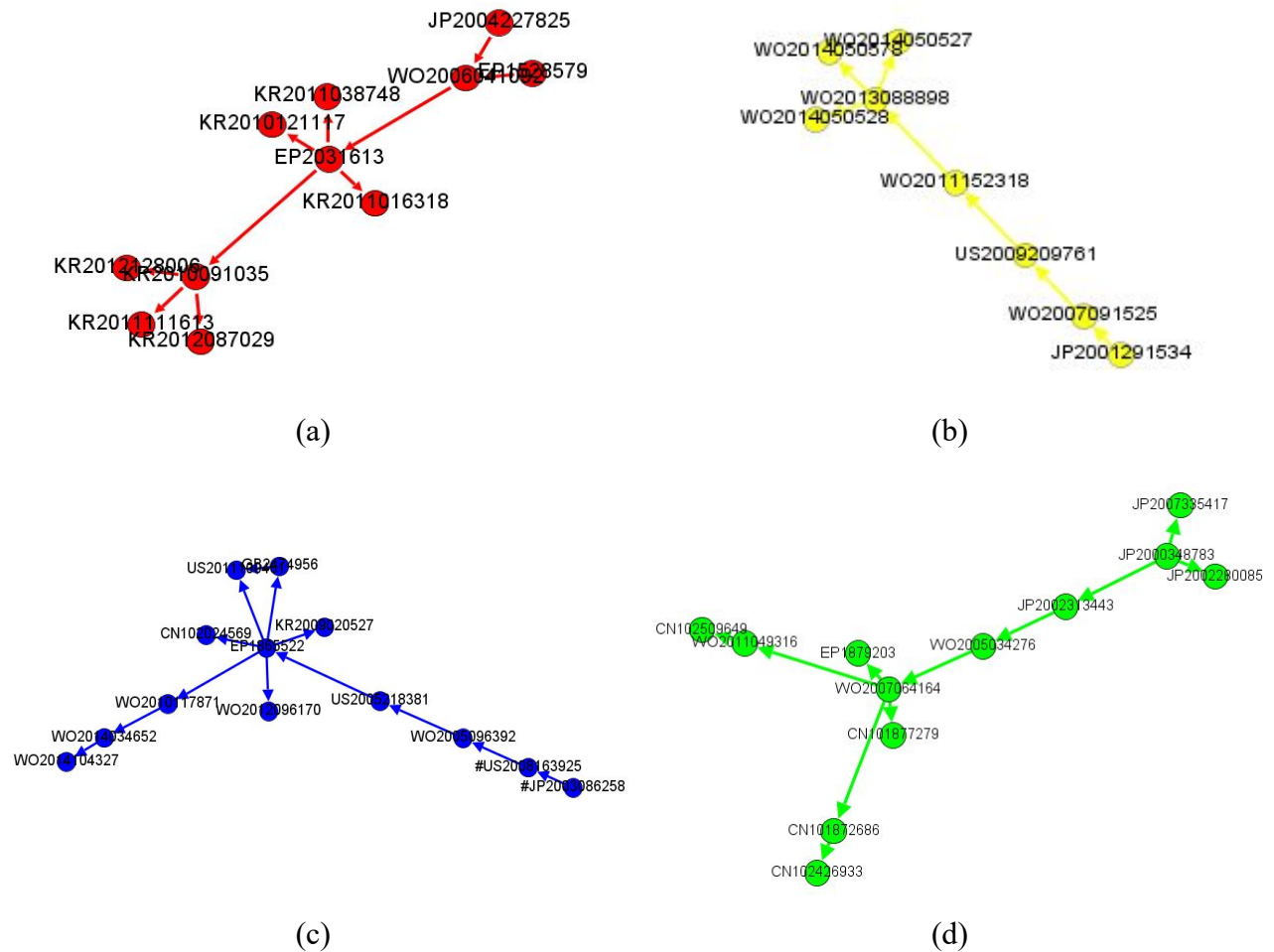


Fig.3 Main technological trajectories of sub-technologies in DSSCs from 1991 to 2014: (a) Photoanode; (b) Sensitiser; (c) Electrolyte and (d) Counter-electrode.

After identifying the patents located on the technological trajectories for the four sub-technologies, we read titles and abstracts, and manually extract the technology focus. Some patents may reflect the same technology focus for their citation relationships. In this case, we prefer to seek the different foci to help researchers and decision-makers identify future research

trends and technology opportunities. These main technological focus are displayed by time period in Figure 4.

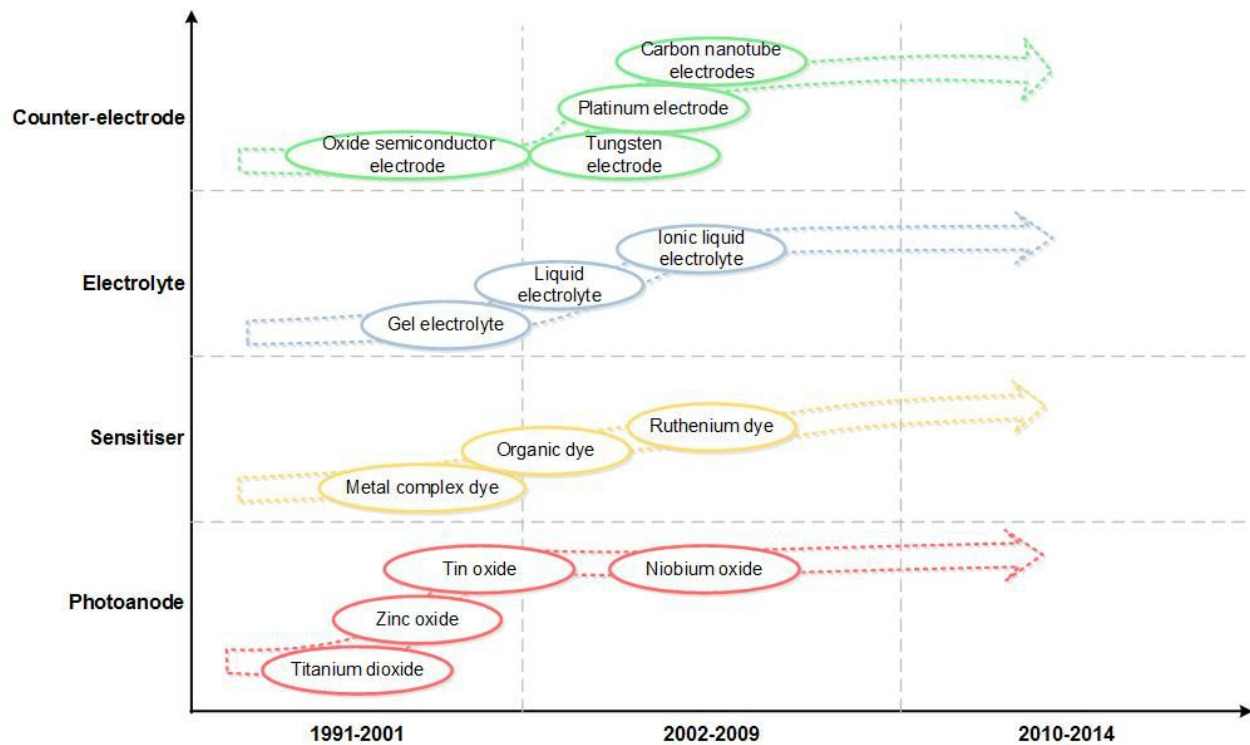


Fig. 4 Preliminary technology evolution pathways of DSSCs (based on the citation analysis).

To further ensure accurate results, we use the technology focus to retrieve the dataset of sub-technologies. Moreover, we ask experts to help evaluate the importance of the technology focus using the "Importance index." Table 4 indicates that most of the technology foci are influential in their actual field, and this evaluation, to some extent, is supported by experts' understanding. Based on such analyses, we have grounds to believe that extracting technology foci from the main citation paths of the patent citations network is both feasible and meaningful.

Table 4. Evaluation of the technology focus included in patents located on main technological trajectories

Sub-technology	Records	Technology focus	Coverage	Importance index
Photoanode	616	Titanium dioxide	28.24%	☆☆☆☆☆
	184	Zinc oxide	8.44%	☆☆☆☆
	116	Tin oxide	5.32%	☆☆☆☆
	22	Niobium oxide	1.01%	☆☆☆
Sensitizer	69	Organic dye	11.22%	☆☆☆☆☆
	34	Metal complex dye	5.53%	☆☆☆☆

	28	Ruthenium complex dye	4.55%	☆☆☆
Electrolyte	108	Liquid electrolyte	6.45%	☆☆☆☆
	104	Ionic liquid electrolyte	6.21%	☆☆☆
	21	Gel electrolyte	1.25%	☆☆☆☆
Counter-Electrode	270	Oxide semiconductor electrode	11.40%	☆☆☆
	146	Platinum electrode	1.39%	☆☆☆☆☆
	77	Carbon electrode	3.25%	☆☆☆☆☆
	68	Tungsten oxide	0.46%	☆☆☆

Note: The number of stars indicates importance. The more stars it earns, the more importance the technology focus presents. Five stars is the highest in our evaluation.

DSSC term clumping analysis

Using the important yet limited information extracted from the trajectory, we use term clumping for each sub-technology to enhance the technology evolution roadmap and to better identify opportunities and possibilities for future trends. The results of the term clumping for DSSCs are shown in Table 5.

Table 5. Term clumping results for DSSCs

<i>Process</i>	<i>All</i>	<i>Photoanode</i>	<i>Sensitiser</i>	<i>Electrolyte</i>	<i>Counter-electrode</i>
Original Phases	75563	33361	11585	24175	28570
Common and Basic Removal	63958	27775	8816	20136	24131
Fuzzy Words Matching	41924	21129	7137	20135	18294
Extreme Words Removal	9238	4864	1683	4696	4534
Combine Term Networks	1562	748	424	411	613

Note: The number shown in the table represents the number of remaining terms after the process

After the term clumping process, we are left with four lists of keywords, from which we choose those that are most related to respective technology foci. Then we extract the first year that these terms appeared in the patent documents. The results are shown in Table 6.

Table 6. Keywords of sub-technologies in DSSCs by term clumping method

<i>Component</i>	<i>Keywords</i>
Photoanode	Indium tin oxide (2002); Silicon oxide (2004); Magnesium oxide (2004); Tungsten oxide (2007); Zirconium oxide (2008); Calcium oxide (2008); Lanthanum oxide (2008); Yttrium oxide (2008); Cerium oxide (2008); Molybdenum oxide (2008); Lithium oxide (2009); Aluminum zinc oxide (2010)
Sensitiser	Polymethine dye (2003); Porphyrin-based dye (2005); Solid state dye (2007); Fluorescent dye (2007); N719 dye (2008); Merocyanine dye (2008); Natural dye

	(2009); Cyanine dye (2009); Phenothiazine dye (2009)
Electrolyte	Non-aqueous electrolyte (1999); Polymer electrolyte (2001); Gel-like electrolyte (2001); Solid electrolyte (2002); Redox electrolyte (2003); Quasi-solid electrolyte (2004); Salt electrolyte (2004)
Counter-electrode	Conductive polymer electrode (2003); Metal electrode (2004)

Based on the preliminary technology evolution (shown in Figure 4), we complement DSSCs sub-technologies keywords to construct the complete technology evolution roadmap (shown in Figure 5). Among these four sub-technologies, the smallest focus is counter-electrode. Platinum counter electrodes, as the most widely used form for counter-electrodes, has the best performance, but its high cost restricts industrialization. Novel materials, such as different nanostructured carbon materials, conductive polymers and their composite counter-electrodes, are drawing attention because of their low cost and high activity. The electrolyte relates closely to the cell's stability, and both counter-electrode and electrolyte haven't achieved a special breakthrough in recent years. In the beginning, the high energy conversion efficiency of DSSCs was achieved with conventional liquid electrolytes, which involve a serious problem of stability. With the development of quasi-solid state and solid state electrolytes, the stability of DSSCs may improve remarkably. Dye sensitizer, which greatly affects the photoelectronic efficiency of solar cells, is an important research focus in the field of cell materials. According to research of the past twenty years, the sensitizers used in DSSCs are mainly divided into two types: metal complex dye and organic dye. Regarding these two types of sensitizers utilized in DSSCs, some different structures and improved dye have been developed. The photoanode plays a role in a cell's performance. Its function is to load sensitizers and collect and transport electrons. Currently, a series of semiconductor materials, including TiO_2 , ZnO , Nb_2O_5 , SnO_2 , etc, are being pursued. We believe that, in the near future, gains in both sensitizers and photoanodes will further improve the efficiency of DSSCs and promote industrialization of these 3rd generation photovoltaic cells.

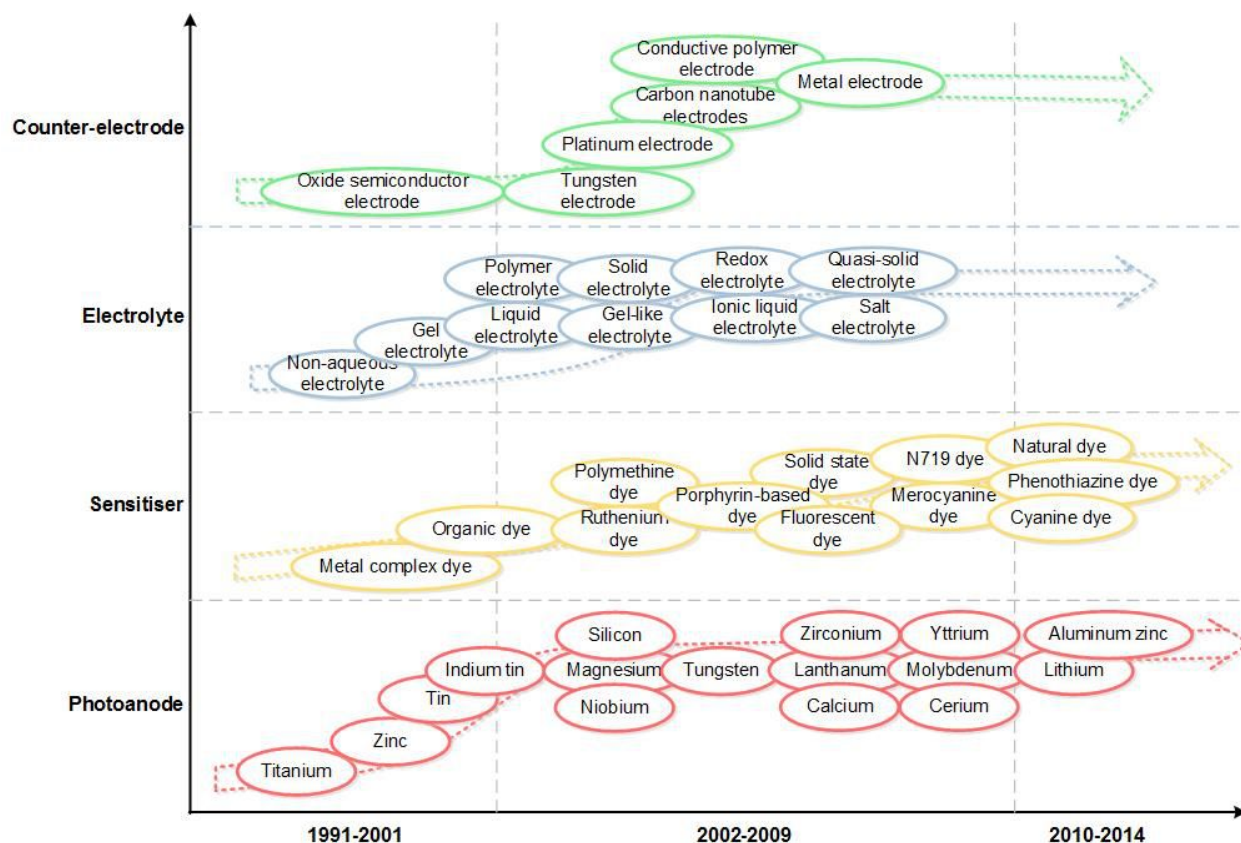


Fig. 5 Final technology evolution pathway of DSSCs

Conclusions

Efficiently generating intelligence on NESTs is an essential topic both for academia and industry. Patents, as a major public resource, offer a wealth of information for conducting technology opportunity analysis. In this paper, we trace technology evolution roadmaps and identify potential low-cost, high efficiency opportunities in the DSSCs field by combining patent citation network analysis with "Tech mining" from the view of non-domain experts.

For a comprehensive technology, it is hard to capture the whole technology evolution at a micro level, so the perspective of sub-technologies becomes necessary. In this paper, with the help of text analytic software, we apply PCA to the top 200 phrases consolidated from selected titles and abstracts. After obtaining the 17 factors and their keywords, we use these keywords and IPCs to extract the sub-technology datasets from the original database. Finally, with the guidance of an expert, we identify four sub-technologies: photoanode, sensitiser, electrolyte and counter-electrode. Overall, this approach helps us to obtain meaningful and useful, albeit incomplete, data in a reasonable way.

Technological hot-topics always play vital roles in the technology development process. Discovering these hot-topics leads to a better understanding of the technology attributes. In this context, we first gain the maximum connected components from the patent citation network of all four datasets, respectively. Then we get the key patents located on the main citation trajectory and manually extract the technological keywords. The results show that, by using this method,

we can obtain essential technology information in the respective fields. This opinion, to some extent, is supported by the judgment of experts. Therefore, extracting the main citation path to obtain the technology focus from the patent citations network is both feasible and meaningful.

"Tech mining", especially term clumping as an effective method of text analysis, helps us to extract terms from titles and abstracts, which provides a abundant basis to further identify highly relevant keywords reflecting the technology focus. We can then go on to improve the accuracy of constructing technology evolution pathways. In this paper, we generated lists of 12, 9, 7, and 2 keywords for photoanode, sensitiser, electrolyte and counter-electrode, respectively, which makes a great contribution to drawing technology evolution pathways for DSSCs.

This paper combines patent citation network analysis with "Tech mining" to trace technology evolution roadmaps, generate a framework of DSSCs development pathways, and discover potential opportunities. The methodology proposed in this paper has some limitations. On the one hand, we combine related keywords and IPCs to obtain a dataset for each sub-technology, but the results are influenced by the keywords we choose and the IPC range we set. Also, different experts and researchers vary in their opinions. On the other hand, though this paper identifies the core hot-topics and important technology focuses, the connection between different technology focuses is hard to grasp without domain specialists.

Some possible future research topics related to this work are as follows: First, publications, as another important source of technological information, can be merged to better trace the technology evolution pathways. Second, more attention should be given to finding reasonable and reliable ways to identify the sub-technologies of a technology field, especially for the NESTs. Third, seeking the relationship and evolution between different technology focuses will be an interesting direction to explore because it is important for policy makers, technology managers and entrepreneurs to better predict the future trends based on recent topics.

Acknowledgments

This research is undertaken at Georgia Institute of Technology, drawing on support from the United States National Science Foundation (NSF) (Award No.1064146), National High Technology Research and Development Program of China (Grant No.2014AA015105) and the China Scholarship Council (CSC Student ID 201406030005). The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the supporters. The authors would like to thank Tingting Ma and Ying Guo for contributions to the DSSCs analyses. We also thank Jan Youtie, Fujin Zhu and other colleagues in the "Innovation Co-lab" of Georgia Institute of Technology, Beijing Institute of Technology, and Manchester University, for their advice and feedback.

References

- [1] Batagelj, V., and Mrvar, A., "Pajek—analysis and visualization of large networks." *Graph Drawing Software*. Jünger, M. and Mutzel, P., Ed. Springer Berlin Heidelberg, 2004.
- [2] Bengisu, M., "Critical and emerging technologies in Materials, Manufacturing, and Industrial Engineering: A study for priority setting," *Scientometrics*, vol. 58, pp. 473-487, 2003.

- [3] Cho, T. S. and Shih, H. Y., "Patent citation network analysis of core and emerging technologies in Taiwan: 1997–2008," *Scientometrics*, vol. 3, pp. 795-811, 2011.
- [4] Choi, C. and Park, Y., "Monitoring the organic structure of technology based on the patent development paths," *Technol. Forecast. Soc. Change*, vol. 76, pp. 754-768, 2009.
- [5] Choi, S., Kim, H., Yoon, J., Kim, K. and Lee, J. Y., "An SAO-based text-mining approach for technology roadmapping using patent information," *R&D Manage.*, vol. 43, pp. 52-74, 2013.
- [6] Dibiaggio, L., and Nesta, L., "Patents statistics, knowledge specialisation and the organisation of competencies," *Revue D'économie Industrielle*, vol. 110, pp.103-126, 2005.
- [7] Érdi, P., Makovi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P., and Zalányi, L., "Prediction of emerging technologies based on analysis of the US patent citation network," *Scientometrics*, vol. 95, pp. 225-242, 2013.
- [8] Harhoff, D., Scherer, F. M., and Vopel, K., "Citations, family size, opposition and the value of patent rights," *Res. Policy*, vol. 32, pp. 1343-1363, 2003.
- [9] Han, K., and Shin, J., "A systematic way of identifying and forecasting technological reverse salients using QFD, bibliometrics, and trend impact analysis: A carbon nanotube biosensor case," *Technovation*. vol. 34, pp. 559-570, 2014.
- [10] Huang, L., Zhang, Y., Guo, Y., Zhu, D., and Porter, A. L., "Four dimensional Science and Technology planning: A new approach based on bibliometrics and technology roadmapping," *Technol. Forecast. Soc. Change*, vol. 81, pp. 39-48, 2014.
- [11] Huang Y., Zhu F. Guo Y., Porter, A.L., and Zhu, D., "Identifying technology evolution pathways based on tech mining and patent citation network- illustrated for dye-sensitized solar cells," *Proceedings-the 5th International Conference on Future-Oriented Technology Analysis (FTA)*. Brussels, Belgium, December 2014.
- [12] Lee, C., Seol, H., and Park, Y., "Identifying new IT-based service concepts based on the technological strength: A text mining and morphology analysis approach," *The 4th International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 4, pp. 36-40, 2007.
- [13] Lee, C. Y., Lee, J. D., and Kim, Y., "Demand forecasting for new technology with a short history in a competitive environment: the case of the home networking market in South Korea," *Technol. Forecast. Soc. Change*, vol. 75, pp. 91-106, 2008.
- [14] Lee, S., Lee, S., Seol, H., and Park, Y., "Using patent information for designing new product and technology: keyword based technology roadmapping," *R&D Manage.*, vol. 38, pp. 169-188, 2008.
- [15] Lee, S., Yoon, B., Lee, C., and Park, J., "Business planning based on technological capabilities: Patent analysis for technology-driven roadmapping," *Technol. Forecast. Soc. Change*, vol. 76, pp. 769-786, 2009.

- [16] Ma, T., Porter, A. L., Guo, Y., Ready, J., Xu, C., and Gao, L., "A technology opportunities analysis model: applied to dye-sensitised solar cells for China," *Technol. Anal. Strateg. Manage.*, vol. 26, pp. 87-104, 2014.
- [17] Milanez, D. H., de Faria, L. I. L., do Amaral, R. M., Leiva, D. R., and Gregolin, J. A. R., "Patents in nanotechnology: an analysis using macro-indicators and forecasting curves," *Scientometrics*, vol. 101, pp. 1097-1112, 2014.
- [18] Michel, J., and Bettels, B., "Patent citation analysis. A closer look at the basic input data from patent search reports," *Scientometrics*, vol. 51, pp. 185-201, 2001.
- [19] Newman, N. C., Porter, A. L., Newman, D., Trumbach, C. C., and Bolan, S. D., "Comparing methods to extract technical content for technological intelligence," *J. Eng. Technol. Manage.*, vol 32, pp. 97-109, 2014.
- [20] O'Brien, J.J., Carley, S., and Porter, A.L., "Keyword field cleaning through ClusterSuite: A termclumping tool for VantagePoint software," *Poster presented at 3rd Global Tech Mining Conference*. Atlanta, USA, September 2013.
- [21] Porter, A. L, and Cunningham, SW., *Tech mining: exploiting new technologies for competitive advantage*. Wiley, New York, 2005.
- [22] Porter, A. L., Guo, Y., and Chiavatta, D., "Tech mining: Text mining and visualization tools, as applied to nanoenhanced solar cells," *Wiley Interdiscip. Rev.-Data Mining Knowl. Discov.*, vol 1, pp. 172-181, 2011.
- [23] Robinson, D. K., Huang, L., Guo, Y., and Porter, A. L., "Forecasting Innovation Pathways (FIP) for new and emerging science and technologies," *Technol. Forecast. Soc. Change*, vol. 80, pp. 267-285, 2013.
- [24] Shibata, N., Kajikawa, Y., Takeda, Y., and Matsushima, K., "Detecting emerging research fronts based on topological measures in citation networks of scientific publications," *Technovation*, vol. 28, pp. 758-775, 2008.
- [25] Thomson Reuters," *DWPI Manual Code Revision.*, "Retrieved 01/21/15 World Wide Web, http://ip-science.thomsonreuters.com/m/pdfs/DWPI_mcr_Jan2015.pdf.
- [26] Watatani, K., Xie, Z., Nakatsuji, N., and Sengoku, S., "Global competencies of regional stem cell research: bibliometrics for investigating and forecasting research trends," *Regen. Med.*, vol. 8, pp. 659-668, 2013.
- [27] Xin, L., Jiwu, W., Lucheng, H., Jiang, L., and Jian, L., "Empirical research on the technology opportunities analysis based on morphology analysis and conjoint analysis," *Foresight*, vol. 12, pp. 66-76, 2010.
- [28] Yoon, B., and Park, Y., "A systematic approach for identifying technology opportunities: Keyword-based morphology analysis," *Technol. Forecast. Soc. Change*, vol. 72, vol. 145-160, 2005.

- [29] Yoon, B., and Park, Y., "Development of new technology forecasting algorithm: Hybrid approach for morphology analysis and conjoint analysis of patent information," *IEEE T. Eng. Manage.*, vol. 54, pp. 588-599, 2007.
- [30] Yoon, B., Phaal, R., and Probert, D., "Morphology analysis for technology roadmapping: application of text mining," *R&D Manage.*, vol. 38, pp. 51-68, 2008.
- [31] Yoon, J., and Kim, K., "An automated method for identifying TRIZ evolution trends from patents," *Expert Syst. Appl.*, vol. 38, pp. 15540-15548, 2011.
- [32] Zhou, X., Zhang, Y., Porter, A. L., Guo, Y., and Zhu, D., "A patent analysis method to trace technology evolutionary pathways," *Scientometrics*, vol. 100, pp. 705-721, 2014.
- [33] Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., and Newman, N. C., ""Term clumping" for technical intelligence: A case study on dye-sensitized solar cells," *Technol. Forecast. Soc. Change*, vol. 85, pp. 26-39, 2014.
- [34] Zhang, Y., Zhou, X., Porter, A. L., and Gomila, J. M. V., "How to combine term clumping and technology roadmapping for newly emerging science and technology competitive intelligence:"problem and solution" pattern based semantic TRIZ tool and case study," *Scientometrics*, vol. 101, pp. 1375-1389, 2014.
- [35] Zhang, Y., Zhou, X., Porter, A. L., Gomila, J. M. V., and Yan, A., "Triple Helix innovation in China's dye-sensitized solar cell industry: hybrid methods with semantic TRIZ and technology roadmapping," *Scientometrics*, vol. 99, pp. 55-75, 2014.