JCDL 2022 COLOGNE

北京物资学院 BEIJING WUZI UNIVERSITY

A hybrid approach to identify and forecast technological opportunities based on topic modeling and sentiment analysis

Tingting Ma（presenter), Ruiping Wang,
Hongshu Chen, Xiao Zhou

3rd Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2022)

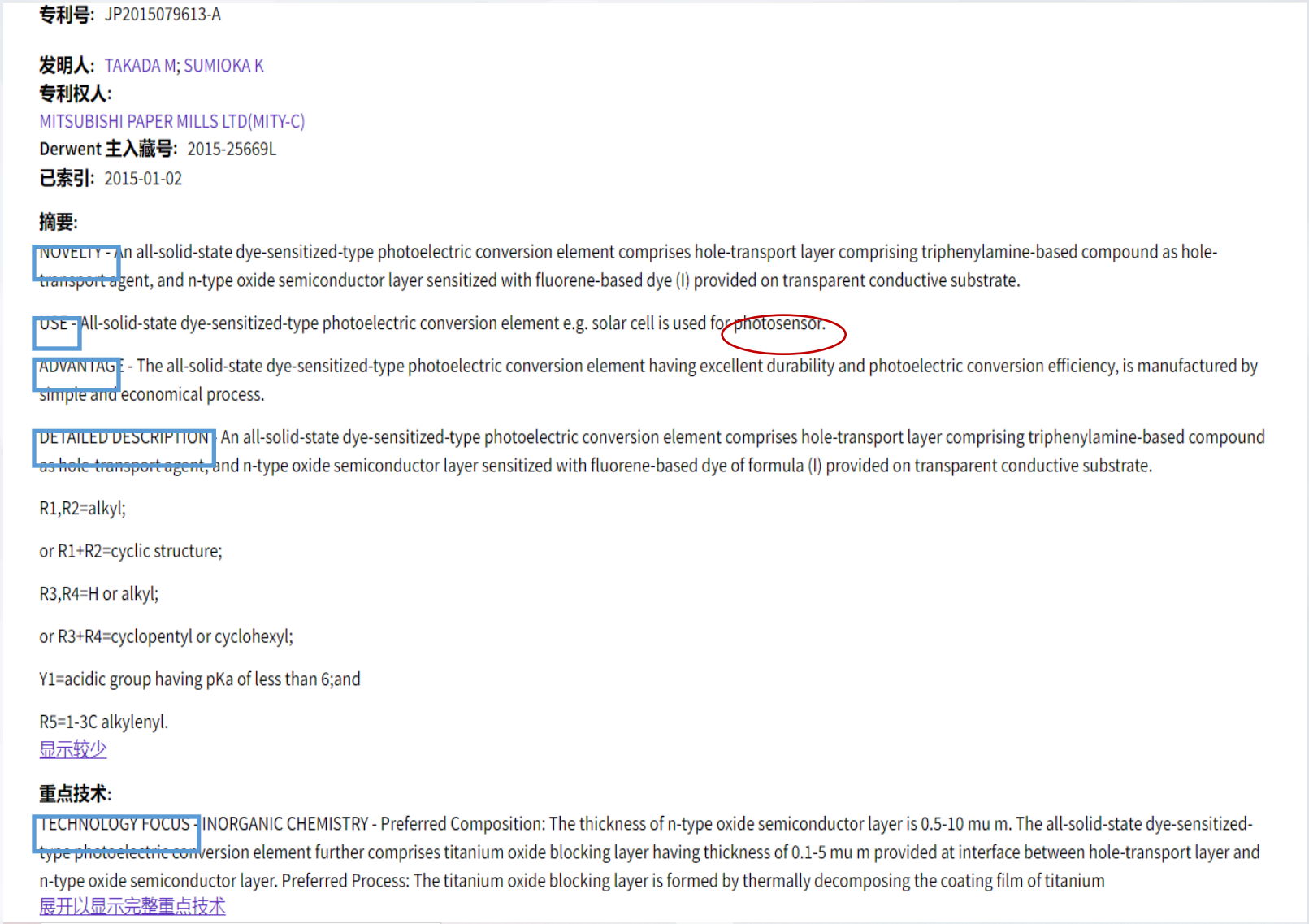
CONTENTS

Background

- Current studies mainly focus on exploring *technical innovative opportunities* by extracting technical topic information from patent document, few studies pays attention to mining *technology application opportunities*.

- Patent document not only records *technical novelties, details* and advantages but also describes *uses* of patented technology.

- **Task1: To identify both technological topics and application topics by mining intelligence separately from the different parts of DII patent document**



专利号: JP2015079613-A

发明人: TAKADA M; SUMIOKA K

专利权人:

MITSUBISHI PAPER MILLS LTD(MITY-C)

Derwent 主入藏号: 2015-25669L

已索引: 2015-01-02

摘要:

NOVELTY - An all-solid-state dye-sensitized-type photoelectric conversion element comprises hole-transport layer comprising triphenylamine-based compound as hole-transport agent, and n-type oxide semiconductor layer sensitized with fluorene-based dye (I) provided on transparent conductive substrate.

USE - All-solid-state dye-sensitized-type photoelectric conversion element e.g. solar cell is used for photosensor.

ADVANTAGE - The all-solid-state dye-sensitized-type photoelectric conversion element having excellent durability and photoelectric conversion efficiency, is manufactured by simple and economical process.

DETAILED DESCRIPTION - An all-solid-state dye-sensitized-type photoelectric conversion element comprises hole-transport layer comprising triphenylamine-based compound as hole-transport agent, and n-type oxide semiconductor layer sensitized with fluorene-based dye of formula (I) provided on transparent conductive substrate.

R1,R2=alkyl;

or R1+R2=cyclic structure;

R3,R4=H or alkyl;

or R3+R4=cyclopentyl or cyclohexyl;

Y1=acidic group having pKa of less than 6;and

R5=1-3C alkylenyl.

显示较少

重点技术:

TECHNOLOGY FOCUS - INORGANIC CHEMISTRY - Preferred Composition: The thickness of n-type oxide semiconductor layer is 0.5-10 mu m. The all-solid-state dye-sensitized-type photoelectric conversion element further comprises titanium oxide blocking layer having thickness of 0.1-5 mu m provided at interface between hole-transport layer and n-type oxide semiconductor layer. Preferred Process: The titanium oxide blocking layer is formed by thermally decomposing the coating film of titanium

展开以显示完整重点技术

Background

- The traditioinal method combining topic modeling and bibliometric analysis has some limitations:

  - ➢ Phrases representing different types of knowledge entities may have diverse effects on topic recognition
  - ➢ Bibliometrics indicators can only quantitatively explain whether the topics are hot or not, but could not show the attitudes of inventors' attitude towards the topics.

➢ Both of **"carbon electrode"** and **"organic dye"** have advantages in **low cost**

➢ "low cost" could be a very common phrases that appears in Advantage parts of DII patents referring to these two technologies.

➢ Task2: Building a term selection model to investigate how advantage terms, use terms and Tech focus terms influence the accuracy of technical topics identification.

Positive attitutes(Examples):

➢ **Advantage**----photoelectric conversion element having excellent durability and photoelectric conversion efficiency

➢ **Novelty**----The counter electrode can be a powerful substitute for a traditional platinum-based counter electrode.

➢ Task3:Integrating sentiment analysis and bibliometric indicators to judge the value of topics from both quantity and quality

Data and Methodology

- **Step1:** Constructing a term selection model, which contained eight combinations of different terms.
- **Step2:** Building a training dataset in which the patents were all labelled categories manually, and created eight "Document-Terms" matrices according to the term selection model.
- **Step3:** Based on the matrices, LDA method were employed to generate latent topics and the topic distributions on patents.
- **Step4:** Assigning each patent to the topic with the highest distribution on it, and the assigning results were compared to the actual categories.
- **Step4:** Calculating the accuracy scores (Total Precision) for the eight combination sets and compared the results of these sets.

$$\text{Total Precision} = \frac{\text{Number of records clustered to correct topic}}{\text{Total Number of records}}$$

TABLE I. TERM SELECTION MODEL

Group	No.	Model Structure	Total precision
A	1#	Title + Abstract	75.92%
	2#	Title + Abstract - Advantage	78.63%
	3#	Title + Abstract - USE	78.09%
	4#	Title + Abstract – USE - Advantage	79.22%
B	1#	Title + Abstract + Tech Focus	76.08%
	2#	Title + Abstract – Advantage + Tech Focus	79.43%
	3#	Title + Abstract – USE + Tech Focus	80.11%
	4#	**Title + Abstract – USE – Advantage + Tech Focus**	**81.50%**

B-#4 term combination, including Tech focus terms but removing Advantage terms and Use terms perform best on technical topic recognition.

Data and Methodology

- A multi-dimensional index system to evaluate technical topics from **quantity** and **quality**

"quantity"

- **Total distribution weight (TDW)** is measured by summing the probabilities of patents distribution on a technical topic to assess the overall attention received by the technology.

$$\text{Total distribution weight (TDW) of Topic k} = \sum_{i=1}^{M} w_{ik}, i = \{1, 2, \cdots, M\}$$

- **Change rate of distribution weight (CRDW)** over time to measure the recent and even near future concern degree of a technology

	Topic1	Topic2	Topic3	L	Topic K	
Document1	0.0027	0.0382	0.0398	L	0.0938	Year1
Document2	0.1903	0.0943	0	L	0.0483	
M	M	M	M	L	M	
Document401	0.0982	0.0763	0.0387	L	0.0273	Year2
Document402	0.0058	0.1983	0.2934	L	0	
M	M	M	M	L	M	
Document851	0	0.0395	0.0964	L	0.0498	Year T
M	M	M	M	L	M	
Document M	0.0049	0	0.0984	L	0.1928	

Fig. 2. An example of a topic distribution matrix in chronological order

$$\text{Change rate of distribution weight(CRDW) of Topic K} = \frac{ADW_k^{2020} + ADW_k^{2019} + ADW_k^{2018}}{ADW_k^{2017} + ADW_k^{2016} + ADW_k^{2015}}$$

"quality"

- Employing the LSTM neural network model to build the classifier to judge the sentiment polarities of all the sentences
- Positive score (PS) of a patent is measured by the number of sentences with positive polarity in the patent.
- we use the patent distribution on a topic to weight the positive scores of its associated patents and calculate the average positive score (APS) to reflect the common judgment of domain experts toward the topic

$$\text{Average positive score (APS) of Topic k} = \frac{\sum_{i=1}^{M} w_{ik} p_i}{\sum_{i=1}^{M} w_{ik}}, \ i = (1, 2, \cdots, M)$$

Case study:DSSC

- Particular technologies with significant potential firstly rise to the surface with high values in most or all indicators. (4 topics)
- Then, by comprehensive analyzing the total distribution weight(TDW) and the positive score (APS), we can find the technologies with low attention but high value, which may be a potential technology opportunity. (3 topics)
- Moreover, combing the change rate of distribution weight (CRDW) and the positive score (APS), the technologies which is hot recently but the value still needs to be improved also can be identified.

Table 5 The indicator value of the 27 topics with the marks

Topic	Indicator value			Mark		
	TDW	CRDW	APS	TDW	CRDW	APS
Organic dye	706	67.0%	2.69	√	√	√
Graphene & Carbon material	354	57.8%	2.56	√	√	√
Organic polymeric material	346	55.9%	2.95	√	√	√
TiO2	340	60.0%	2.48	√	√	√
Apparatus or power supply system containing DSSC	711	88.7%	2.44	√	√	
Photoanode modified method	520	88.4%	2.36	√	√	
Photoelectric conversion layer	677	34.6%	2.61	√		√
Light absorption layer	421	50.9%	2.64	√		√
Metal oxide semiconductor layer	392	33.3%	2.47	√		√
Glass type sealing material	386	44.2%	2.52	√		√
Metal catalyst	385	46.6%	2.85	√		√
Polymer electrolyte	497	55.4%	2.46	√		
Structure of solar cell	439	46.9%	2.23	√		
Metal substrate	410	35.9%	2.30	√		
Organic hole transport materials	179	98.2%	2.75		√	√
P-type/n-type semiconductor	154	74.7%	2.51		√	√
Preparation of nano materials	155	68.7%	2.35		√	
Sulfide for counter electrode	141	88.9%	2.44		√	
Electrode active metal material	135	50.7%	2.52			√
Conductive polymer	319	42.7%	2.26			
Laminated solar-cell module	284	42.3%	2.42			
Transparent conductive film	250	37.8%	2.45			
Organic solvent electrolyte	215	46.7%	2.35			
Metal complex dye	214	39.2%	2.21			
Lonic liquid electrolyte	204	51.8%	2.32			
ZnO	187	45.4%	2.28			
Metal oxide semiconductor particles	169	47.1%	2.38			
Average Scores of Indicator	**340**	**55.5%**	**2.47**			

Case study:DSSC

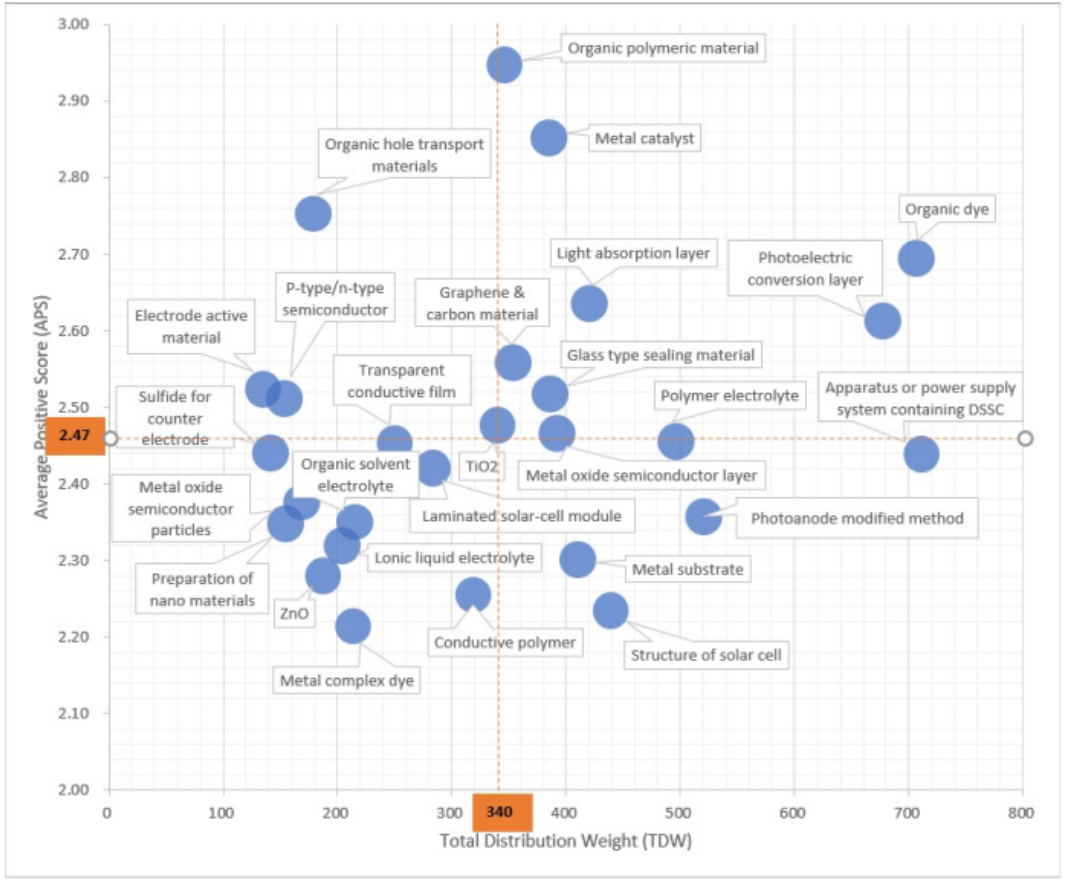


Figure 3 The TDW-APS two-dimensional scatter diagram for the technical topics of DSSCs

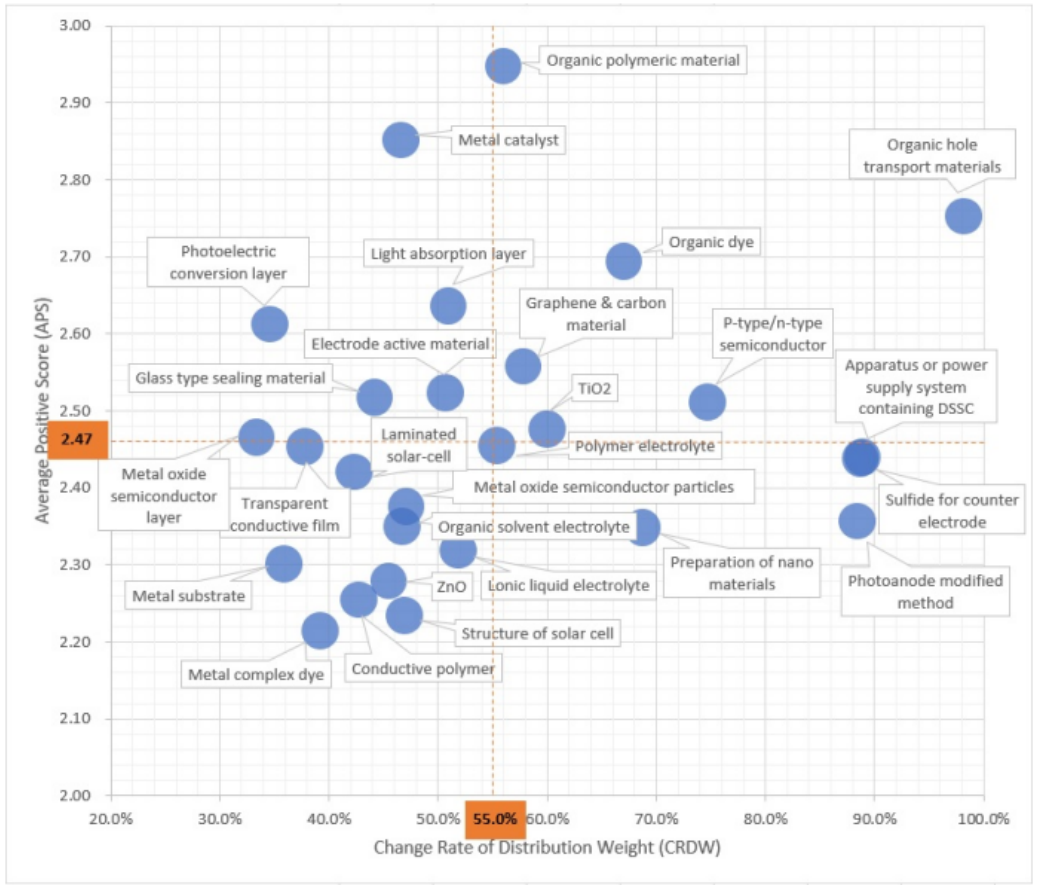
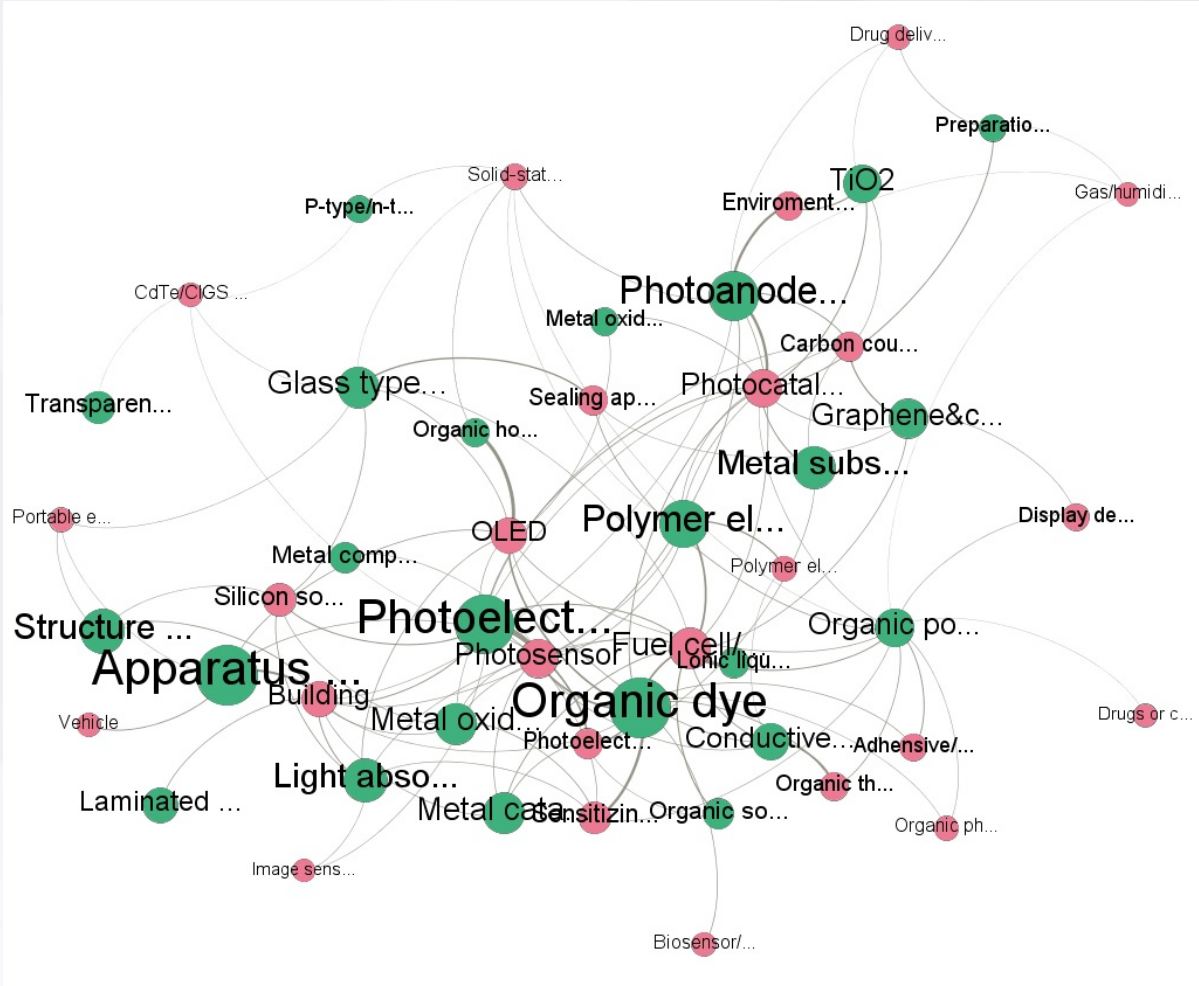


Figure 4 The CRDW-APS two-dimensional scatter diagram for the technical topics of DSSCs

Case study:DSSC

Table 7 The topics on the applications of DSSC sub-technologies

No.	Topics	TDW	No.	Topics	TDW
1	Fuel cell/lithium ion batter	385	14	Display device	149
2	Photosensor	343	15	Solid-state DSSC	133
3	Photocatalyst film	332	16	Polymer electrolyte	130
4	OLED	298	17	Organic photoelectric conversion element	113
5	Building	295	18	Drug delivery system	107
6	Silicon solar cell	244	19	Biosensor/electrochemical sensor	105
7	Sensitizing dye	239	20	CdTe/CIGS solar cell	105
8	Photoelectric transducer	205	21	Vehicle	103
9	Sealing application	198	22	Portable electronic product	102
10	Carbon counter electrode	192	23	Gas/humidity/molecule sensor	98
11	Environment purification	187	24	Drugs or cosmetics	85
12	Organic thin-film solar cell	174	25	Image sensors	84
13	Adhensive/coating/packaging material	159			



We proposed a probability-based topic relation measurement method to link the identified technologies with their highly related technical problems and applications, in order to figure out advancing ways and promising applications for specific technologies.

Conclusion

- There are three main contributions of this study.
  - First, we distinguished the terms in the different parts of patent documents when utilizing them to recognize topics, and created a term selection model to measure and compare the different terms' influences on topic recognition. The analysis results tell us that the Tech focus terms contributes to topic recognition, while the Advantage and the Use part have much noisy terms which may obfuscate technical topic identification.
  - Second, we identified both technological topics and application topics by mining different parts of Derwent patent.
  - Third, we introduce sentiment analysis to support topic assessment and construct a multi-dimensional evaluation system to identify potential innovation opportunities from both the "quantity" and "quality" aspects.

Conclusion

- There are also some parts should be improved in the further.
  - First, the performance of the LDA model on recognizing topics from short texts should be validated. Thus, in next step we can employ several candidate topic modeling methods to select the optimal one.
  - Second, we simply use positions of phrases and their semantic structure to judge entity types of phrases. This method is too rough. In the follow-up, we will try to label the entity type of the phrase by using deep learning technologies.