Applying IPC-Based Clustering and Link Analysis to Patent Analysis on Thin-Film Solar Cell

Tzu-Fu Chiu

Department of Industrial Management and Enterprise Information
Aletheia University, Taiwan
chiu@mail.au.edu.tw

Abstract. Patent analysis has been recognized as an important task at the government and company levels. Patent data contain plentiful technical information, which is worthwhile to be used in patent analysis in order to find out the technical categories and the technological trend. Due to the complex nature of patent data, two data mining methods: IPC-based clustering and link analysis, are used to figure out the possible technological trend on thin-film solar cell. IPC-based clustering, a proposed clustering method for exploiting the professional knowledge of the patent office examiners, will be utilized to generate the IPC-based clusters via the IPC and Abstract fields; while the link analysis will be adopted to draw a link diagram via the Abstract, Issue Date, and Assignee Country fields. During experiment, the major technical categories will be identified using IPC-based clustering, and the technological trend will be recognized through the link diagram. Finally, the major technical categories and technological trend will be provided to the managers and stakeholders for assisting their decision making.

Keywords: IPC-based clustering, link analysis, patent analysis, thin-film solar cell.

1 Introduction

Solar cell (especially thin-film solar cell) is a key technology option to realize the shift to a decarbonized energy supply and tends to offer a reduction of prices, rather than an increase in the future [1]. In addition, up to 80% of the technological information disclosed in patents is never published in any other form [2]. Meanwhile, patent analysis has been recognized as an important task at the government and company levels. Through appropriate analysis, technological details and relations, business trends, novel industrial solutions, or investment policy making can be achieved [3]. Due to the textual characters of patent data (in Abstract, Claim, and Description fields), a clustering method, IPC-based clustering is proposed to manipulate the homogeneity and heterogeneity of patents on the Abstract field so as to increase the cohesiveness (similarity) within a cluster and the dispersion (dissimilarity) among clusters. In patents, the IPC codes are provided by the examiners of patent office and contain the professional knowledge of the examiners

N.T. Nguyen (Ed.): Transactions on CCI XII, LNCS 8240, pp. 108–122, 2013.

[4]. It would be reasonable for a research to base on the term vectors of Abstract and the IPC codes to classify the patents into a certain number of clusters for facilitating patent analysis. Afterward link analysis is employed to find out the relations between year (/country) and cluster. Consequently, the technical categories will be obtained and the technological trend of thin-film solar cell will be recognized for assisting the decision making of managers and stakeholders.

2 Related Work

As this study is attempted to observe the technological trend for companies and stakeholders via patent data, a research framework is required and can be built via a utilization of IPC-based clustering (a modified clustering method) and an adoption of link analysis. In order to manipulate the homogeneity and heterogeneity of patent data, IPC-based clustering is proposed for dividing the patents into different clusters. Due to the collected data spreading over ten years (2000 to 2009) and in different countries, link analysis is employed to generate the linkages between year (/country) and clusters. Subsequently, the research framework will be applied to identify the technical categories and to recognize the technological trend on thin-film solar cell. Therefore, the related areas of this study would be patent analysis and technological trend, patent data and thin-film solar cell, IPC-based clustering, and link analysis, which will be described briefly in the following subsections.

2.1 Patent Analysis and Technological Trend

Patent analysis has been reviewed in the literature [5-8] and can be classified as: country level (policy making and international comparison), industry level (science and technology, knowledge spillovers, and competitive intelligence), organization level (technology licensing, corporate strategy, and business function), and technology level (technology development and product management) [8]. This study, attempting to explore the technological trend, is in the type of industry level (competitive intelligence).

Trend analysis is the practice of collecting information and attempting to spot a pattern, or trend, in the information [9]. Technology forecasting is to predict a moving trend of technological change. It also supports mining knowledge for technology marketing and reducing risk of R&D investment in company and government [10]. Moreover, technological trend investigation is useful for finding promising business fields in the future and for detecting the direction of competitive technical development, for examples: the trend of market entry, the trend of technological evolution, and the maturity of fields (matured, maturing, or undeveloped) [11]. Consequently, in order to draw the data mining techniques for observing the technological trend via patent data, a research framework will be designed by using IPC-based clustering and link analysis to perform the patent analysis on thin-film solar cell in this study.

2.2 Patent Data and Thin-Film Solar Cell

A patent document is similar to a general document, but includes rich and varied technical information as well as important research results [3]. Patent data, among the better structured and monitored data sources, is the official filings of inventions [12]. Patent documents can be gathered from a variety of sources, such as the United States Patent and Trademark Office [13], the European Patent Office [14], the Intellectual Property Office in Taiwan [15], and so on. A patent document includes numerous fields [13], such as: Patent Number, Title, Abstract, Issue Date, Application Date, Application Type, Assignee Name, Assignee Country, International Classification (IPC), Current US References, Claims, Description, etc.

Photovoltaics (PV) is the technology that generates direct current electrical power from semiconductors (or some other materials) when they are illuminated by photons [16]. Solar cell is the basic building block of solar photovoltaics and a sort of green energy. It can be mainly divided into two categories (according to the light absorbing material): crystalline silicon (in a wafer form) and thin films (of other materials) [17]. A thin-film solar cell (TFSC), also called a thin-film photovoltaic cell (TFPV), is made by depositing one or more thin layers (i.e., thin film) of photovoltaic material on a substrate [18]. The most common materials of TFSC are amorphous silicon and polycrystalline materials (such as: CdTe, CIS, and CIGS) [17]. In 2009, the photovoltaic industry production increased by more than 50% (yearly growth rates in average over the last decade: 40%) and reached a world-wide production volume of 11.5 GWp of photovoltaic modules, whereas the thin film segment grew faster than the overall PV market [1]. Therefore, thin film is the most potential segment with the highest production growth rate in the solar cell industry, and it would be appropriate for academic and practical researchers to contribute efforts to explore the technological trend of this segment.

2.3 IPC-Based Clustering

An IPC (International Patent Classification) is a classification derived from the International Patent Classification System (supported by WIPO) which provides a hierarchical system of symbols for the classification of patents according to the different areas of technology to which they pertain [19]. IPC classifies technological fields into five hierarchical levels: section, class, subclass, main group and sub-group, containing about 70,000 categories [20]. As stated by the Intellectual Property Office of UK [4], each patent document published will have at least one IPC code applied to it; and the EPO and other patent offices worldwide also use it to classify their own patent documents. The IPC codes of every patent are assigned by the examiners of the national patent office and contain the professional knowledge of the experienced examiners [21]. Therefore, it would be reasonable for a research to base on the IPC code and the term vectors of Abstract to cluster the patents into a number of categories. The IPC codes have been applied for assisting patent retrieval in some researches [21, 22].

As IPC codes of patent are provided by the examiners and contain professional knowledge, they are suitable to be exploited to direct the clustering process. So, a modified clustering method, IPC-based clustering, is proposed to include the IPC codes to enrich the clustering mechanism in this study. The idea of this method is also based on the author's previous studies [23, 24]. However, some differences between this study and the previous ones are: the IPC-based clustering method was modified and rewritten in more detail and more precisely; the research framework was reconstructed to be more appropriate for guiding the experiment; and the paper was reorganized so as to state the problem domain, the related work, the problem solving approach, and the experiment and explanation more clearly and completely.

The processes of the IPC-based clustering are IPC Group Centroid Generation, IPC-based Cluster Generation, Clustering Alternative Generation, and Optimal Alternative Selection, which are explained as follows:

(1) IPC Group Centroid Generation: The patents with the same IPC code will be put together to form an IPC code group (G_i), if a patent which has more than one IPC code will be assigned to multiple groups. Patents in the same IPC code group will then be applied to calculate a group centroid (called IPC group centroid) c_i using the term vector of the Abstract field (i.e., x_{ij}) via Equation (1) where G_i is the ith group.

$$c_i = \frac{1}{|G_i|} \sum_{x \in G_i} x_{ij} \tag{1}$$

(2) IPC-Based Cluster Generation: According to the IPC group centroids and term vectors, the whole dataset of patents will be distributed into a certain number of clusters via the Euclidean distance measure as in Equation (2) [25] where x_{ij} is a term vector of patent in G_i.

$$d(x_{ij}, c_i) = \sqrt{(x_{ij} - c_i)^2} \tag{2}$$

A patent will be assigned to a specific IPC code cluster according to the shortest distance $d(x_{ij}, c_i)$ existing between that patent and the IPC code centroid c_i. The patents distributed to a code group form an IPC-based cluster.

(3) Clustering Alternative Generation: The first clustering alternative is made initially by including its composing IPC-based clusters of 4 clusters. The following alternatives are then made successively by 5 clusters up to a certain number (e.g., 31 in this study), which is determined based on the research requirements and the domain knowledge. Furthermore, for enhancing the accuracy of clustering, an adjusted method is suggested which retains the larger clusters (with more patents) from a potential alternative by setting the threshold of the number of comprising patents to a suitable value (e.g., 6 in this study) and then repeat the "IPC-based Cluster Generation" again to obtain an adjusted alternative. Subsequently, the original and adjusted alternatives will be used to form the overall clustering alternatives.

(4) Optimal Alternative Selection: Among the clustering alternatives, F score (in Equation (3)) is employed to evaluate the accuracy of the clustering results, where the Precision and Recall are in Equation (4) and (5) [26].

$$F = \frac{2}{(1/\mathrm{Re}call) + (1/\mathrm{Pr}ecision)} \tag{3}$$

$$\mathrm{Pr}ecision = \frac{\left|\{relevant \cap retrieved\}\right|}{\left|\{retreieved\}\right|} \tag{4}$$

$$\mathrm{Re}call = \frac{\left|\{relevant \cap retrieved\}\right|}{\left|\{relevant\}\right|} \tag{5}$$

An original or adjusted alternative with the higher F score will be selected as an optimal alternative of the clustering result.

2.4 Link Analysis

Link analysis is a collection of techniques that operate on data that can be represented as nodes and links [27]. A node represents an entity such as a person, a document, or a bank account. A link represents a relationship between two entities such as a parent/child relationship between two people, a reference relationship between two documents, or a transaction between two bank accounts. The focus of link analysis is to analyze the relationships between entities. The areas related to link analysis are: social network analysis, search engines, viral marketing, law enforcement, and fraud detection [27]. In search engines, the page rank of page A, $PR(A)$, can be calculated as in Equation (6), where T_j is a page pointing to A; $C(T_j)$ is the number of going out links from page T; and d is a minimum value assigned to any page [28, 29]. In social network analysis, the degree centrality of a node can be measured as in Equation (7), where $a(P_i, P_k) = 1$ if and only if P_i and P_k are connected by a link (0 otherwise) and n is the number of all nodes [30]. Additionally, in data mining, the relationship strength (i.e., the similarity between nodes) can be measured by Jaccard coefficient as in Equation (8), where r_i is the ith record of a data set [31, 32].

$$PR(A) = d + (1-d) * \sum_{j=1}^{n} (PR(T_j) / C(T_j)) \tag{6}$$

$$\mathrm{C}_D(P_k) = \sum_{i=1}^{n} a(P_i, P_k) / (n-1) \tag{7}$$

$$Ja(r_i, r_j) = \frac{Freq(r_i \cap r_j)}{Freq(r_i \cup r_j)} \tag{8}$$

In this study, link analysis will be employed to generate the linkages between the year (/country) and the IPC-based cluster for the technological trend observation.

3 A Research Framework for Technological Trend Observation

A research framework for the technological trend observation, based on IPC-based clustering and link analysis, has been constructed as shown in Fig. 1. It consists of

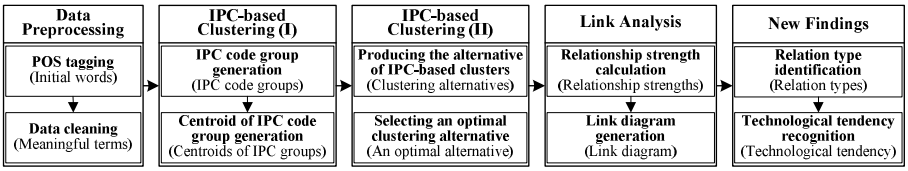

Fig. 1. A research framework for the technological trend observation

five phases: data preparation, IPC-based clustering (I), IPC-based clustering (II), link analysis, and new findings; and will be described in the following subsections.

3.1 Data Preprocessing

In first phase, the patent data of thin-film solar cell (during a certain period of time) will be downloaded from the USPTO [13]. For considering an essential part to represent a patent document, the Abstract, Issue Date, and Assignee Country fields are selected as the objects for this study. Afterward, two processes, POS tagging and data cleaning, will be executed to clean up the textual data of the abstract field.

(1) POS Tagging: An English POS tagger (i.e., a Part-Of-Speech tagger for English) from the Stanford Natural Language Processing Group [33] will be employed to perform word segmenting and labeling on the patents (i.e., the abstract field). Then, a list of proper morphological features of words needs to be decided for sifting out the initial words.

(2) Data Cleaning: Upon these initial words, files of n-grams, synonyms, and stop words will be built so as to combine relevant words into compound terms, to aggregate synonymous words, and to eliminate less meaningful words. Consequently, the meaningful terms will be obtained from this process.

3.2 IPC-Based Clustering (I)

Second phase is intended to describe the first two steps of the IPC-based clustering as stated in Subsection 2.3. The IPC code group generation is used to generate the IPC code groups according to the IPC field of patent data. The centroid of IPC code group generation is utilized to calculate the centroids of IPC code groups based on the patents in every group.

(1) IPC code group generation: A patent is distributed to an IPC code group if the patent contains that specific IPC code. Since a patent holds at least one to several IPC codes, a patent will be distributed to one or to several IPC code groups. For example, Patent 07605328 contains H01L031/00, B05D005/12, and H01L02/00 codes; and will be distributed to these three IPC code groups.

(2) Centroid of IPC code group generation: The comprising patents of an IPC code group are used to calculate the centroids for that IPC code group according to

Equation (1). For example, 48 patents of the first IPC code group (G_1) will be used to calculate its centroid (c_1).

3.3 IPC-Based Clustering (II)

Third phase is utilized to depict the other two steps of the IPC-based clustering. Producing the alternative of IPC-based clusters is applied to produce successively a series of clustering alternatives, each one consisting of a certain number of clusters. Selecting an optimal clustering alternative is to select an optimal alternative based on the F score measure for generating the technical categories.

(1) Producing the Alternative of IPC-Based Clusters: A clustering alternative is made by including its composing IPC-based clusters (e.g., 4 clusters in the first alternative; and 5 to 31 clusters in the following alternatives). An adjusted alternative is obtained by retaining the larger clusters (with more patents) from a potential alternative via setting the threshold of the number of comprising patents to a suitable value (e.g., 6 in this study) and then redistributing the patents again to the retained clusters. Both the original and adjusted alternatives form the clustering alternatives

(2) Selecting an Optimal Clustering Alternative: Among the clustering alternatives, F score (in Equation (3)) is employed to evaluate the accuracy of the clustering results. An original or adjusted alternative with the higher F score will be selected as the appropriate clustering result. In the selected alternative, every IPC-based cluster is regarded as a technical category and will be utilized for further analysis in the next phase.

3.4 Link Analysis

Third phase is designed to perform the link analysis for producing the relationship strengths and the link diagram so as to obtain the relations between years (/countries) and technical categories.

(1) Relationship Strength Calculation: In order to generate the summary table and link diagram, the relationship strength between nodes and the centrality of nodes (i.e., technical categories, years, and countries) need to be calculated via Equation (5) and Equation (4) respectively. The calculation result of relationship strength will be summarized in tables, so as to facilitate the identification of the linkages between year (/countries) and technical categories.

(2) Link Diagram Generation: Based on the summary tables and the node centrality, a link diagram will be drawn, so that the relations between year (/countries) and technical categories can be constructed through the threshold settings of relationship strength and node centrality. These relations will be utilized to identify the relation types between the year (/countries) and technical categories and then to explore the technological trend in the following phase.

3.5 New Findings

Last phase is intended to identify the relation types between technical categories and years (/countries) and to recognize the technological trend, based on the relationship strengths and the link diagram.

(1) Relation Type Identification: According to the relationship strengths and the link diagram, the relation types between the technical categories and years (/countries) will be identified. For the relations between categories and years, four relation types are likely found: a category existing in the full period of time (i.e., not less than 5) (type A1), existing in the first half (type A2), existing in the second half (type A3), and existing randomly in the period of time (type A4). For the relations between the categories and countries, three relation types are likely found: a category spreading in various countries (i.e., not less than 5) (type B1), spreading in the dominant countries (i.e., JP with 70 patents and US with 52 patents) (type B2), and spreading in the non-dominant countries (type B3).

(2) Technological Trend Recognition: In accordance with the relation types of technical category, the technological trend of thin-film solar cell will be recognized and then provided to the managers and stakeholders for assisting their decision making.

4 Experimental Results and Explanation

The experiment has been implemented according to the research framework. The experimental results will be explained in the following five subsections: result of data preprocessing, result of IPC code group and IPC group centroid, result of clustering alternatives and IPC-based clusters, result of link analysis, and result of new findings.

4.1 Result of Data Preprocessing

As the aim of this study is to explore the trends of thin-film solar cell, the patent documents are the target data for the experiment. Mainly, the Abstract, IPC, Issue Date, and Country fields were used in this study. The issued patents (160 records) during year 2000 to 2009 were collected from USPTO (USPTO, 2010), using key words: "'thin film' and ('solar cell' or 'solar cells' or 'photovoltaic cell' or 'photovoltaic cells' or 'PV cell' or 'PV cells')" on "title field or abstract field". Afterward, the POS tagger was triggered and the data cleaning process was executed to do the data preprocessing upon the Abstract data. Consequently, the Abstract data during year 2000 to 2009 were cleaned up and the meaningful terms were obtained.

4.2 Result of IPC Code Group and IPC Group Centroid

According to the IPC field, the number of IPC code groups (down to the fifth level) in 160 patents were 190, as many patents contained more than one IPC code, for

example, Patent 06420643 even contained 14 codes. But there were up to 115 groups consisting of only one patent. If the threshold of the number of comprising patents was set to 5, there were 31 leading groups including the first group H01L031/18 (consisting of 48 patents), the second group H01L021/02 (27 patents), and till to the 31st group H01L031/0236 (5 patents), as in Fig. 2.

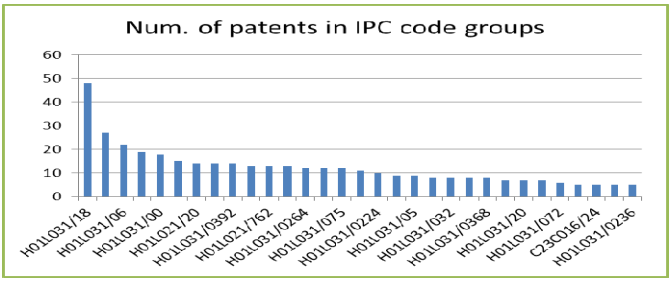

Fig. 2. The number of patents in IPC code groups

The patents contained in each IPC code group were used to generate the IPC group centroid for that group via Equation (1). These IPC group centroids would be utilized to produce the IPC-based clusters afterward.

4.3 Result of Clustering Alternatives and IPC-Based Clusters

The clustering alternatives contained the ones from including 4 leading clusters, to 5 leading clusters, …, till 31 leading clusters. The first alternative was constructed by distributing patents of the whole dataset into the 4 leading clusters (i.e., H01L031/18,

Table 1. A summary of clustering alternatives with their including clusters and accuracies

Alternative	Num. of clusters	Num. of patents in the IPC-based cluster	Accuracy
1	4	118, 11, 19, 12	0.4514
2	5	93, 11, 15, 10, 31	0.5109
3	6	87, 11, 15, 10, 29, 8	0.5213
4	7	84, 5, 15, 8, 29, 8, 11	0.5140
5	8	77, 5, 12, 8, 26, 8, 11, 13	0.5055
6	9	77, 5, 12, 1, 26, 8, 11, 13, 7	0.5052
7	10	73, 5, 12, 1, 24, 8, 11, 10, 7, 9	0.4988
8	15	55, 2, 2, 1, 24, 8, 0, 7, 10, 9, 10, 0, 10, 10, 12	0.5466
9	20	48, 2, 2, 0, 16, 8, 0, 5, 8, 5, 9, 0, 10, 10, 11, 8, 1, 6, 8, 3	0.5251
10	25	47, 2, 2, 0, 16, 7, 0, 5, 6, 5, 5, 0, 5, 10, 1, 8, 0, 6, 7, 3, 10, 2, 2, 5, 6	0.4992
11	31	42, 2, 0, 0, 16, 7, 0, 2, 6, 5, 5, 0, 4, 7, 1, 8, 0, 3, 6, 3, 10, 2, 0, 0, 6, 5, 4, 5, 5, 2, 4	0.5192
adjusted	9	70, 23, 15, 12, 10, 10, 7, 7, 6	0.5617

H01L021/02, H01L031/06, and H01L031/036) via Equation (2), using the IPC group centroids of 4 clusters and the term vectors of Abstract data. The number of patents distributed into 4 clusters was: 118 in H01L031/18, 11 in H01L021/02, 19 in H01L031/06, and 12 in H01L031/036. The accuracy (i.e., average *F* score) of this alternative was 0.4514. Each IPC group with its distributed patents was regarded as an IPC-based cluster. The other alternatives (from including 5 to 31 clusters) were then constructed successively. Some of the clustering alternatives with their including IPC-based clusters and accuracies were calculated and summarized as in Table 1.

According to Table 1, the accuracies of most alternatives varied from 0.50 to 0.53. The adjusted method was applied to pinpoint the leading clusters from the potential alternative 11 (with 31 clusters) by setting the threshold of the number of comprising patents to 6, so as to increase the accuracy of clustering to 0.5617. After the IPC-based Cluster Generation, the adjusted alternative, including 9 IPC-based clusters: H01L031/18, H01L031/00, H01L021/00, H01L021/20, H01L031/052, H01L031/048, H01L031/04, H01L031/0336, and H01L031/20, was the appropriate alternative as shown in Table 2 (with the containing patents and IPC code description). These nine IPC-based clusters were regarded as the major technical categories and used in the following link analysis.

Table 2. The appropriate alternative with including IPC-based clusters and IPC code description

IPC-based cluster	Num. of patents	IPC code description
H01L031/18	70	Processes or apparatus specially adapted for the manufacture or treatment of these devices or of parts thereof
H01L031/00	23	Semiconductor devices sensitive to infra-red radiation, light, electromagnetic radiation of shorter wavelength, or corpuscular radiation and specially adapted either for the conversion of the energy of such radiation into electrical energy or for the control of electrical energy by such radiation; Processes or apparatus specially adapted for the manufacture or treatment thereof or of parts thereof; Details thereof
H01L031/048	15	encapsulated or with housing
H01L031/052	12	with cooling, light-reflecting or light-concentrating means
H01L021/20	10	Deposition of semiconductor materials on a substrate, e.g. epitaxial growth
H01L031/0336	10	in different semiconductor regions, e.g. Cu_2X/CdX hetero-junctions, X being an element of the sixth group of the Periodic System
H01L021/00	7	Processes or apparatus specially adapted for the manufacture or treatment of semiconductor or solid state devices or of parts thereof
H01L031/04	7	adapted as conversion devices
H01L031/20	6	such devices or parts thereof comprising amorphous semiconductor material

4.4 Result of Link Analysis

Using link analysis, the relationship strengths between the major technical categories and years (/countries) were calculated and summarized in Table 3 and 4, where the items in italic face were the ones not less than the threshold setting: 0.05.

Table 3. The relationship strengths between major technical categories and years

Category	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
H01L031/18	***0.0588***	***0.1325***	***0.0824***	***0.1646***	***0.0741***	0.0274	***0.0921***	***0.1579***	0.0411	***0.0506***
H01L031/00	0.0238	0.0444	***0.0976***	0.0465	***0.0526***	0	0.0286	***0.1081***	***0.0741***	***0.1613***
H01L031/048	***0.0606***	***0.1143***	***0.1563***	0.0278	***0.0667***	0	0.0370	0	0	0
H01L031/052	0	***0.1250***	***0.0968***	***0.0625***	0.0357	0	0.0417	0.0345	0	0
H01L021/20	***0.1538***	0.0303	0.0323	***0.0667***	0.0385	***0.0714***	0	0	0	0
H01L031/0336	***0.2000***	0	***0.0667***	0	0.0385	***0.0714***	0	0	0	0.0455
H01L021/00	0	0	0	0	0.0435	***0.0909***	***0.0526***	0.0417	***0.0833***	***0.1111***
H01L031/04	0.0385	0	0	0.0357	***0.1429***	0	***0.0526***	0	0	***0.0526***
H01L031/20	***0.0833***	***0.0714***	0	0.0370	0	0	***0.0556***	0	0	0

Table 4. The relationship strengths between major technical categories and countries

Category	JP	US	DE	NL	FR	KR	AU	CA	BE	IT	CH	TH	FI	TW
H01L031/18	***0.2500***	***0.2323***	***0.1467***	0.0411	0.0139	*0*	*0*	0.0286	0.0141	*0*	*0*	0.0143	*0*	*0*
H01L031/00	***0.0690***	***0.2500***	0.0263	*0*	*0*	*0*	*0*	*0*	*0*	*0*	0.0435	*0*	*0*	*0*
H01L031/048	***0.1333***	0.0308	*0*	***0.0500***	***0.0588***	*0*	*0*	*0*	*0*	*0*	*0*	*0*	***0.0667***	*0*
H01L031/052	***0.0513***	0.0492	***0.0769***	*0*	*0*	*0*	***0.1667***	*0*	***0.0769***	*0*	*0*	*0*	*0*	*0*
H01L021/20	***0.1111***	0.0333	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
H01L031/0336	***0.0667***	0.0333	0.0400	***0.0667***	***0.0833***	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
H01L021/00	0.0132	***0.0727***	*0*	*0*	*0*	***0.1250***	*0*	*0*	*0*	***0.1429***	*0*	*0*	*0*	*0*
H01L031/04	***0.0694***	*0*	*0*	*0*	*0*	***0.1250***	*0*	*0*	*0*	*0*	*0*	*0*	*0*	***0.1429***
H01L031/20	0.0411	0.0175	0.0476	***0.0909***	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*

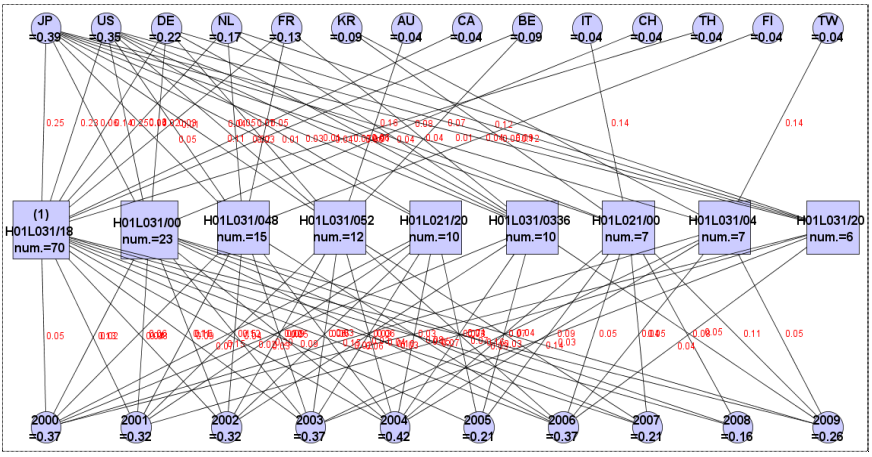

Fig. 3. A link diagram for 9 major technical categories (H01L031/18 to H01L031/20)

Based on Table 3 and 4, a link diagram for 9 major technical categories was drawn via the relationship strengths (not less than the threshold 0.05) between the categories and the years (/countries) and the centralities of the year and country nodes in order to demonstrate the relations between the categories and the years (/countries), which was

shown in Fig. 3. In the diagram, the digits under the year and country nodes were the centralities (e.g., year 2000: 0.37; JP: 0.39); the digits on the link lines were the relationship strengths (e.g., between H01L031/18 and 2000: 0.05; between H01L031/18 and JP: 0.25).

4.5 Result of New Findings

The link diagram (i.e., Fig. 3) would be utilized to identify the relation types. Afterward, relation types would be applied to explore the technological trend.

(1) Relation Type Identification: According to the link diagram (Fig. 3) and the summarized table (Table 3), the relation type A1 (between technical categories and years) were: H01L031/18 and H01L031/00. The relation type A2 were: H01L031/048, H01L031/052, H01L021/20, and H01L031/0336. The relation type A3 were: H01L021/00 and H01L031/04. The relation type A4 was: H01L031/20. In addition, according to the link diagram (Fig. 3) and the summarized table (Table 4), the relation type B1 (between technical categories and countries) were: H01L031/18, H01L031/048, H01L031/052, and H01L031/0336. The relation type B2 were: H01L031/00, H01L021/20, and H01L031/20. The relation type B3 were: H01L021/00 and H01L031/04. Subsequently, the relation types between major technical categories and years as well as between major technical categories and countries were summarized below in Table 5 and then used to recognize the technological trend.

Table 5. A summary of major technical categories and relation types

Category	Focused year	Type	Related country	Type
H01L031/18	2000, 2001, 2002, 2003, 2004, 2006, 2007, 2009	A1	JP, US, DE, NL, FR,CA, BE, TH	B1
H01L031/00	2002, 2004, 2007, 2008, 2009	A1	JP, US, DE, CH	B2
H01L031/048	2000, 2001, 2002, 2004	A2	JP, US, NL, FR, FI	B1
H01L031/052	2001, 2002, 2003	A2	JP, US, DE, AU, BE	B1
H01L021/20	2000, 2003, 2005	A2	JP, US	B2
H01L031/0336	2000, 2002, 2005	A2	JP, US, DE, NL, FR	B1
H01L021/00	2005, 2006, 2008, 2009	A3	JP, US, KR, IT	B3
H01L031/04	2004, 2006, 2009	A3	JP, KR, TW	B3
H01L031/20	2000, 2001, 2006	A4	JP, US, DE, NL	B2

(2) Technological Trend Observation: According to the link diagram (Fig. 3) and the above summarized table (Table 5), the technological trend of thin-film solar cell could be observed. As then major technical categories were: H01L031/18, H01L031/00, H01L021/00, H01L021/20, H01L031/052, H01L031/048, H01L031/04, H01L031/0336, and H01L031/20, the technological trend of each technical category would be observed and described as follows.

(a) H01L031/18 category: Referring to Table 5, this technical category was continuously developing in the full period of time from 2000 to 2009 (relation type A1) and widely spreading in eight countries (relation type B1). It seemed to be an essential category of the industry, as it is related to the “manufacturing processes or devices”.

(b) H01L031/00 category: From the above Table 5, this category existed in the whole period of time from 2002 to 2009 (type A1) and spread in the dominant countries (type B2). It seemed that the category was growing constantly and participated by the technologically advanced countries. It is related to "semiconductor devices sensitive to infra-red radiation" and "the conversion of the energy".

(c) H01L031/048 category: According to Table 5, this category existed in the first half of the period of time (type A2) and spread in the various countries (type B1). It was likely that the category had been active during 2000 to 2004 and was out of the technical mainstream afterward. It was emphasized by several countries. It is about the "encapsulated or with housing".

(d) H01L031/052 category: Referring to Table 5, this category existed in the first half of the period of time (type A2) and spread in the various countries (type B1). It seemed that the category had been popular during 2001 to 2003 and declined gradually. It is relating to the "cooling, light-reflecting or light-concentrating means".

(e) H01L021/20 category: From the above Table 5, this category existed in the first half of the period of time (type A2) and spread in the dominant countries (type B2). It was likely that the category had been common during 2000 to 2005 and became minor afterward. It was focused mainly by the dominant countries JP (Japan) and US (United States). It is regarding the "deposition of semiconductor materials on a substrate".

(f) H01L031/0336 category: According to Table 5, this category existed in the first half of the period of time (type A2) and spread in the various countries (type B1). It was plausible that the category had been active during 2000 to 2005 and became unimportant eventually. It was stressed by several countries. It is concerning the "different semiconductor regions, e.g. Cu2X/CdX hetero-junctions".

(g) H01L021/00 category: Referring to Table 5, this category existed in the second half of the period of time (type A3) and spread in the non-dominant countries (type B3). It seemed that the category became popular lately from 2005 to 2009 and was contributed by the non-dominant countries as KR (Korea) and IT (Italy). It is relating to the manufacturing processes or devices of semiconductor.

(h) H01L031/04 category: From the above Table 5, this category existed in the second half of the period of time (type A3) and spread in the non-dominant countries (type B3). It was likely that the category gained emphasis slowly from 2004 to 2009 and was participated by the non-dominant countries like KR (Korea) and TW (Taiwan). It is concerning the "adapted as conversion devices".

(i) H01L031/20 category: According to Table 5, this category existed randomly in the period of time (type A4) and spread in the dominant countries (type B2). It seemed that the category was not in the technical mainstream and supported randomly the dominant countries as JP, US, DE (Germany) and NL (Netherlands). It is regarding the "devices comprising amorphous semiconductor material".

In addition, the significant H01L031/18 category possesses 70 patents (about 44%), which reflects that a big portion of patents put efforts in the manufacturing process and device aspects. The dominant countries JP and US possess 70 and 52 patents (about 44% and 32%) respectively, which shows that these two countries held the powerful innovative ability and resources in this industry and can affect the technological trend strongly.

5 Conclusions

The research framework (based on IPC-based clustering and link analysis) for observing the technological trend on thin-film solar cell has been formed. The experiment was performed and the experimental results were obtained. The major technical categories were: H01L031/18, H01L031/00, H01L021/00, H01L021/20, H01L031/052, H01L031/048, H01L031/04, H01L031/0336, and H01L031/20. The technological trend was as follows. The specific categories which existed in the full period of time and developed continuously were: H01L031/18 and H01L031/00 categories. The specific categories which existed in the first half of the period of time and became active earlier were: H01L021/00, H01L021/20, H01L031/052, and H01L031/048 categories. The specific categories which existed in the second half of the period of time and became common lately were: H01L031/04 and H01L031/0336 categories. The dominant countries which possessed the powerful innovative ability and resources in the thin-film solar cell industry were Japan and United States. The above experimental results and findings would be helpful to the managers and stakeholders for their decision making on R&D aspects.

In the future work, the other aspects of company information (e.g., the public announcement, open product information, and financial reports) can be included so as to enhance the validity of research result. Additionally, the patent database can be expanded from USPTO to WIPO or TIPO in order to perform the technological trend observation on thin-film solar cell widely.

Acknowledgments. This research was supported by the National Science Council of the Republic of China under the Grants NSC 99-2410-H-156-014.

References

1. Jager-Waldau, A.: PV Status Report 2010: Research, Solar Cell Production and Market Implementation of Photovoltaics, JRC Scientific and Technical Reports (2010)
2. Blackman, M.: Provision of Patent Information: A National Patent Office Perspective. World Patent Information 17(2), 115–123 (1995)
3. Tseng, Y., Lin, C., Lin, Y.: Text Mining Techniques for Patent Analysis. Information Processing and Management 43, 1216–1247 (2007)
4. Intellectual Property Office, Patent classifications (March 15, 2011), http://www.ipo.gov.uk/pro-types/pro-patent/p-class.htm
5. Basberg, B.L.: Patents and the Measurement of Technological Change: A Survey of the Literature. Research Policy 16, 131–141 (1987)
6. Ashton, W.B., Sen, R.K.: Using Patent Information in Technology Business Planning - I. Research Technology Management 31(6), 42–46 (1988)
7. Breitzman, A.F., Mogee, M.E.: The Many Applications of Patent Analysis. Journal of Information Science 28(3), 187–205 (2002)
8. Lai, K.K., Lin, M.L., Chang, S.M.: Research Trends on Patent Analysis: An Analysis of the Research Published in Library's Electronic Database. The Journal of American Academy of Business 8(2), 248–253 (2006)
9. Wikipedia, Trend analysis (March 16, 2012), http://en.wikipedia.org/wiki/Trend_analysis
10. Jun, S.: A Forecasting Model for Technological Trend using Unsupervised Learning. In: Kim, T.-h., Adeli, H., Cuzzocrea, A., Arslan, T., Zhang, Y., Ma, J., Chung, K.-i., Mariyam, S., Song, X. (eds.) DTA / BSBT 2011. CCIS, vol. 258, pp. 51–60. Springer, Heidelberg (2011)

11. Willfort, Technological Trend Investigation (March 16, 2012), http://www.willfort.com/english2/index.html
12. Russell, S.: Technology Forecasting. In: Narayanan, V.K., O'Connor (eds.) Encyclopedia of Technology and Innovation Management, pp. 37–45. John Wiley & Sons (2010)
13. USPTO (2010) USPTO: the United States Patent and Trademark Office (July 14, 2010), http://www.uspto.gov/
14. EPO (2010) EPO: the European Patent Office (July 14, 2010), http://www.epo.org/
15. TIPO (2010) TIPO: the Intellectual Property Office (July 14, 2010), http://www.tipo.gov.tw/
16. Luque, A., Hegedus, S.: Handbook of Photovoltaic Science and Engineering. John Wiley and Sons (2003)
17. Solarbuzz, Solar Cell Technologies (October 20, 2010), http://www.solarbuzz.com/technologies.htm
18. Wikipedia, Thin film solar cell (October 20, 2010), http://en.wikipedia.org/wiki/Thin_film_solar_cell
19. WIPO, Preface to the International Patent Classification (IPC) (October 30, 2010), http://www.wipo.int/classifications/ipc/en/general/preface.html
20. Sakata, J., Suzuki, K., Hosoya, J.: The Analysis of Research and Development Efficiency in Japanese Companies in the Field of Fuel Cells using Patent Data. R&D Management 39(3), 291–304 (2009)
21. Kang, I.S., Na, S.H., Kim, J., Lee, J.H.: Cluster-based Patent Retrieval. Information Processing & Management 43(5), 1173–1182 (2007)
22. Chen, Y.L., Chiu, Y.T.: An IPC-based Vector Space Model for Patent Retrieval. Information Processing & Management 47(3), 309–322 (2011)
23. Chiu, T.F., Hong, C.F., Chiu, Y.T.: To Propose Strategic Suggestions for Companies via IPC Classification and Association Analysis. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part I. LNCS, vol. 6591, pp. 218–227. Springer, Heidelberg (2011a)
24. Chiu, T.-F., Hong, C.-F., Chiu, Y.-T.: Using IPC-based Clustering and Link Analysis to Observe the Technological Directions. In: Katarzyniak, R., Chiu, T.-F., Hong, C.-F., Nguyen, N.T., et al. (eds.) Semantic Methods for Knowledge Management and Communication. SCI, vol. 381, pp. 183–197. Springer, Heidelberg (2011b)
25. Feldman, R., Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press (2007)
26. Hotho, A., Nürnberger, A., Paaß, G.: A Brief Survey of Text Mining. LDV Forum - GLDV Journal for Language Technology and Computational Linguistics 20(1), 19–62 (2005)
27. Donoho, S.: Link Analysis. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, 2nd edn., pp. 355–368. Springer (2010)
28. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems 30, 107–117 (1998)
29. Weiss, S.M., Indurkhya, N., Zhang, T.: Fundamentals of Predictive Text Mining. Springer (2010)
30. Freeman, L.C.: Centrality in Social Networks: Conceptual Clarification. Social Networks 1, 215–239 (1979)
31. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson Addison-Wesley (2006)
32. Wikipedia, Jaccard index (March 5, 2012), http://en.wikipedia.org/wiki/Jaccard_index
33. Stanford Natural Language Processing Group, Stanford Log-linear Part-Of-Speech Tagger (October 15, 2009), http://nlp.stanford.edu/software/tagger.shtml