# To Propose Strategic Suggestions for Companies via IPC Classification and Association Analysis

Tzu-Fu Chiu[1], Chao-Fu Hong[2], and Yu-Ting Chiu[3]

[1] Department of Industrial Management and Enterprise Information, Aletheia University, Taiwan, R.O.C.
chiu@mail.au.edu.tw
[2] Department of Information Management, Aletheia University, Taiwan, R.O.C.
cfhong@mail.au.edu.tw
[3] Department of Information Management, National Central University, Taiwan, R.O.C.
gloria@mgt.ncu.edu.tw

**Abstract.** Strategic suggestions are essential for companies to facilitate the top management to foresee and review the future directions of their company's research and development investment. Therefore, a research design has been formed for performing the strategic planning on technology where IPC classification was employed to divide the patents into different categories and association analysis was adopted to discover the relations between terms and between clusters. Consequently, the visualized results, crystallized diagrams and integrated map, were generated and the relations between technical topics and companies were observed. Finally, according to the relations, the strategic suggestions on thin-film solar cell for companies were recognized and proposed.

**Keywords:** strategic suggestion, IPC classification, association analysis, thin-film solar cell, patent data.

## 1 Introduction

Solar cell, one of green energies, is growing at a fast pace with its long-lasting and non-polluting natures. To understand the situation and trend of this technology, especially the thin-film solar cell, is essential for companies to foresee and review the future directions of their research and development activities. In technological information, up to 80% of the disclosures in patents are never published in any other form [1]. Additionally, patent analysis has been recognized as an important task at the company, industry, and government levels [2]. Apart from those existing analysis methods such as task identification, searching segmentation, abstracting, clustering, and visualization [2], a research design of IPC classification and association analysis for conducting strategic planning will be built for patent analysis in order to propose the strategic suggestions for companies in the thin-film solar cell industry.

## 2 Related Work

As this study is attempted to propose strategic suggestions for the industry and companies via patent data, a research design is required and can be built via a consideration of IPC

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 218–227, 2011.

classification and association analysis. In order to manipulate the homogeneity and heterogeneity of patent data, IPC classification is employed to divide the patents into different categories. For handling the textual nature of patent data (mainly the abstract, claim, and description fields), association analysis (including data crystallization) is adopted to discover the relations between terms and between clusters. Subsequently, the research design will be applied to the domain of strategic planning on thin-film solar cell. Therefore, the related areas of this study would be strategic planning, thin-film solar cell, IPC classification, and association analysis, which will be described briefly in the following subsections.

### 2.1 Strategic Planning

Strategic planning (also called strategic management) is an organization's process of defining its strategy, or direction, and making decisions on allocating its resources to pursue this strategy, including its capital and people [3]. Strategic planning processes are: mission definition, objectives setting, external analysis, internal analysis, strategic choice, strategy implementation, and competitive advantages [4]. Among them, the strategic choice (i.e., strategy formulation), following the decision-making process, is to develop and evaluate strategic alternatives and then select strategies that support and complement each other and that allow the organization to best capitalize on its strengths and environmental opportunities [5]. In order to draw the data mining techniques for conducting the strategy formulation via patent data, a research design of IPC classification and association analysis will be formed to propose the strategic suggestions for companies on thin-film solar cell in this study.

### 2.2 Thin-Film Solar Cell

Solar cell, a sort of green energy, is clean, renewable, and good for protecting our environment. 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) [6]. 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 [7]. The most common materials of TFSC are amorphous silicon or polycrystalline materials (such as: CdTe, CIS, and CIGS) [6]. In recent years (2003-2007), total PV production grew in average by almost 50% worldwide, whereas the thin film segment grew in average by over 80% and reached 400 MW or 10% of total PV production in 2007 [8]. 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 and propose strategic suggestions for this technology.

### 2.3 IPC Classification

Classification, or categorization, is to classify a given data instance into a prespecified set of categories [9]. The classification task can be defined as to approximate an unknown category assignment function $F : D \times C \rightarrow \{0,1\}$, where $D$ is the set of all possible

documents and $C$ is the set of predefined categories [9]. IPC (International Patent Classification) provides a hierarchical system of symbols for the classification of patents according to the different areas of technology to which they pertain [10]. IPC classifies technological fields into five hierarchical levels: section, class, subclass, main group and sub-group, containing 70,000 categories [11]. In this study, IPC code will be used to divide the patents into different categories.

### 2.4 Association Analysis

Association analysis is a useful method for discovering interesting relationships hidden in large data sets. The uncovered relationships can be represented in the form of association rules or co-occurrence graphs [12]. An event map, a sort of co-occurrence graphs, is a two-dimension undirected graph, which consists of event clusters, visible events, and chances [13]. An event cluster is a group of frequent and strongly related events. The occurrence frequency of events and co-occurrence between events within an event cluster are both high. The co-occurrence between two events is measured by the Jaccard coefficient as in Equation (1), where $e_i$ is the $i$th event in a data record (of the data set $D$). The event map is also called as an association diagram in this study.

$$Ja(e_i, e_j) = \frac{Freq(e_i \cap e_j)}{Freq(e_i \cup e_j)} \tag{1}$$

Secondly, data crystallization is employed to add extra data elements into the association diagram for observing the relations between clusters. Data crystallization is a technique to detect the unobservable (but significant) events via inserting these unobservable events as dummy items into the given data set [14]. The unobservable events and their relations with other events are visualized by applying the event map. A generic data crystallization algorithm can be summarized as follows [15]: (a) event identification, (b) clustering, (c) dummy event insertion, (d) co-occurrence calculation, and (e) topology analysis. The co-occurrence between a dummy event and clusters is measured by equation (2), where $DE_i$ is a dummy event and $C$ is the specific number of clusters.

$$Co(DE_i, C) = \sum_{j=0}^{|C|-1} \max_{e_k \in C_j} Ja(DE_i, e_k) \tag{2}$$

Data crystallization was originally proposed to deal with unobservable events (i.e., dummy events) so as to emerge the hidden clues from existing circumstances via judging the unknown relations [14]. This method has been modified by the authors to insert extra data elements (e.g., patent-no, assignee, or country fields) as dummy events into the original data set (i.e., the abstract field), so that the relations between the extra data elements and existing clusters can come out and be observed [16].

In this study, the association analysis and modified data crystallization will be adopted to generate association diagrams and crystallized diagrams for strategic planning.

## 3 A Research Design for Strategic Planning

As this study is attempted to propose the strategic suggestions for thin-film solar cell, a research design, based on the IPC classification and association analysis, has been developed and shown in Fig. 1. It consists of four phases: data preprocessing, patent classification, association analysis, and new findings; and will be described in the following subsections.

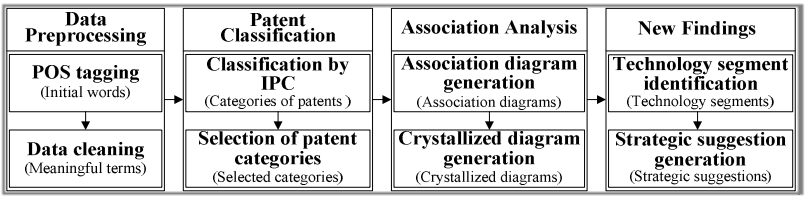


**Fig. 1.** A research design for strategic planning

### 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 [17]. For considering an essential part to represent a patent document, the abstract, assignee, and 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 [18] 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 Patent Classification

Second phase is used to conduct the patent classification according to the IPC field. The patents of thin-film solar cell will be classified by IPC code and selected via the number of appearing times of every specific IPC code.

**(1) Classification by IPC:** Every patent will be assigned to a specific category according to its IPC code and may be assigned to several categories if its IPC field contains more than one code. Meanwhile, the number of patents for an IPC category will be counted for showing its occurrence frequency.

**(2) Selection of Patent Categories:** Based on a threshold setting (e.g., 15 was set for this study) on the number of patents in an IPC category, a certain number of the

leading IPC categories will be selected and then be utilized to generate the association diagrams and crystallized diagrams.

### 3.3 Association Analysis

Third phase is designed to perform the association analysis on the meaningful terms of the abstract data for every IPC category using association diagram generation and crystallized diagram generation so as to obtain the technical topics and relations between topics and companies.

**(1) Association Diagram Generation:** An association diagram will be drawn via the term frequency and co-occurrence from the meaningful terms (of the abstract data) of every IPC category, so that a number of clusters will be generated through the proper thresholds setting of frequency and co-occurrence. These clusters are regarded as technical topics and will be named using the domain knowledge.

**(2) Crystallized Diagram Generation:** In order to generate a crystallized diagram, a dummy event (i.e., assignee) needs to be inserted into the abstract data. Afterward, using the updated abstract data, the diagram will be drawn to show the clusters, dummy nodes, and links. Secondly, the clusters in crystallized diagrams will also be named and regarded as in the association diagrams. Thirdly, the topics, dummy nodes, and links will be utilized to observe the relations between topics and companies. Finally, these crystallized diagrams will be utilized to form an integrated map in the next phase.

### 3.4 New Findings

Last phase is intended to identify the technology segments and to recognize the strategic suggestions, based on the crystallized diagrams of IPC categories.

**(1) Technology Segment Identification:** By combining the crystallized diagrams, an integrated map of technical topics for IPC categories will be constructed via linking the technical topics to the same clusters in the different crystallized diagrams. Subsequently, the technology segments will be identified based on the integrated map.

**(2) Strategic Suggestion Generation:** According to the above integrated map and technology segments, the relations between technical topics and companies can be observed and the strategic suggestions can be recognized. The strategic suggestions of thin-film solar would be useful for the top management to foresee and review the future directions of their company's research and development activities.

## 4 Experimental Results and Explanation

The experiment has been implemented according to the research design. The experimental results will be explained in the following four subsections: result of data preprocessing, result of patent classification, result of association analysis, and result of new findings.

### 4.1 Result of Data Preprocessing

As the aim of this study is to propose the strategic suggestions, the patents of thin-film solar cell are the target data for the experiment. Mainly, the abstract, assignee, and country fields were used in this study. The issued patents (160 records) during year 2000 to 2009 were collected from USPTO, 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 abstract data. Consequently, the abstract data during year 2000 to 2009 were cleaned up and the meaningful terms were obtained.

### 4.2 Result of Patent Classification

According to the IPC field, the number of IPC categories (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 categories individually consisting of only one patent. The leading frequent 50 categories were illustrated in Fig. 2, where the first category H01L031-18 consisted of 48 patents and the 49$^{th}$ category C03B033-07 consisted of 2 patents. By setting the threshold of the number of consisting patents to 15, the leading six IPC categories: H01L031-18 (consisting 48 patents), H01L021-02 (27 patents), H01L031-06 (22 patents), H01L031-036 (19 patents), H01L031-00 (18 patents), and H01L021-00 (15 patents), were selected and would be used to generate the association diagrams, crystallized diagrams, as well as the integrated map. Consequently, the strategic suggestions would be drawn up based on these visualized results.

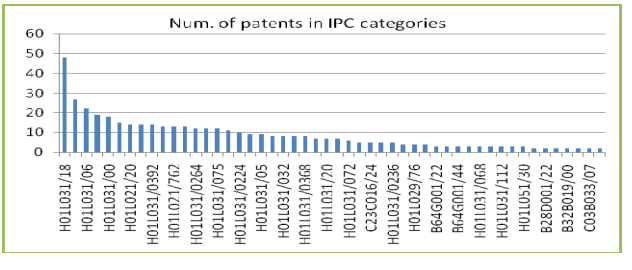


**Fig. 2.** The number of patents in IPC categories

### 4.3 Result of Association Analysis

Using the meaningful terms of abstract data (H01L031-18 to H01L021-00), six association diagrams were drawn via 'association diagram generation' successively. Taking the diagram of H01L031-18 as an example, twelve clusters were found while the number of consisting nodes of a cluster was set to no less than four. According to the domain knowledge, the clusters were named as follows: a1-porous-structure, a2-plasma-CVD, a3-accumulated-charge, a4-amorphous-annealing-deposition, a5-encapsulant, a6-absorber-layer, a7-reflective-film, a8-silicon-film, a9-composite-structure, a10-annealing-process, a11-semiconductor-layer, and a12-compound-semiconductor. These

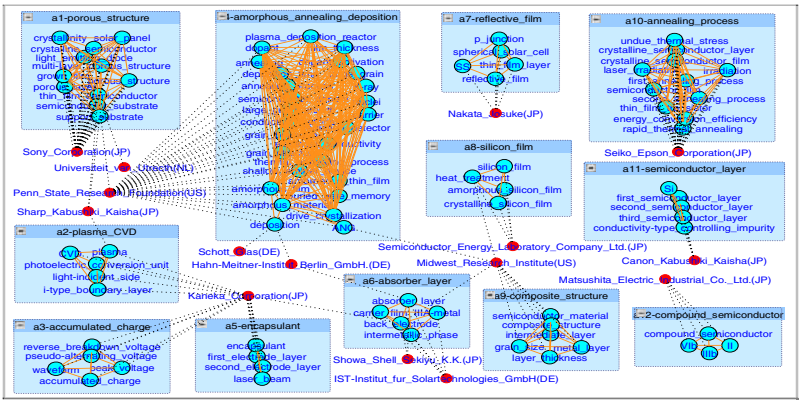


**Fig. 3.** A crystallized diagram of IPC category "H01L031-18"

named clusters were regarded as technical topics. Afterward, using the same abstract data with inserted dummy event (i.e., company-name), six crystallized diagrams were drawn via 'crystallized diagram generation', showing the relations between topics and companies. Fig. 3 is an example of crystallized diagram of IPC category "H01L031-18".

By combining six 'crystallized diagram of IPC category' (H01L031-18 to H01L021-00), an integrated map of technical topics for IPC categories (in Fig. 4) was constructed via linking every technical topic to the same clusters in the six categories.

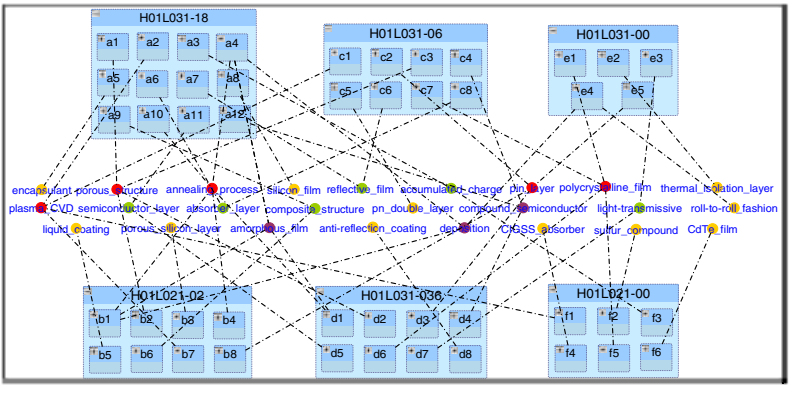


**Fig. 4.** An integrated map of technical topics for IPC categories (H01L031-18 to H01L021-00)

## 4.4 Result of New Findings

The integrated map of technical topics for IPC categories will be utilized to identify the technology segments. Afterward, the integrated map and technology segments will be applied to recognize the possible strategic suggestions.

**(1) Technology Segment Identification:** According to the above integrated map, the relations among technology segments, technical topics, and related companies were

summarized below in Table 1. Depending on the number of clusters which a technical topic was comprised, the technology segments were identified as: significant segment (linking to equal or greater than 4 clusters; the nodes were in red color); regular segment (equal to 3 clusters; in purple color); minor segment (equal to 2 clusters; in green color); and niche segment (equal to 1 cluster; in yellow color). This integrated map and the above crystallized diagrams were then used to recognize the possible strategic suggestions.

**Table 1.** Technology segments with their technical topics and related companies

| Technology segments | Technical topics | H01L0 31-18 | H01L0 21-02 | H01L0 31-06 | H01L0 31-036 | H01L0 31-00 | H01L0 21-00 | Related companies |
|---|---|---|---|---|---|---|---|---|
| Significant segment | porous-structure | a1 | b2 | c3 | d2 | | | Sony(JP), U_Utrecth(NL), Penn_State(US), Sharp(JP), National_Institute(JP), Canon(JP) |
| | annealing-process | a4, a10 | b1, b4 | | d1 | | | Schott(DE), Hahn-Meitner(DE), Seiko_Epson(JP), Penn_State(US), Matsushita(JP) |
| | plasma-CVD | a2 | b7 | c1 | | | f1 | Sharp(JP), Kaneka(JP), Matsushita(JP) |
| | pin-layer | | b8 | c2 | d3 | e5 | | Fuji_Electric(JP), Sharp(JP), Kaneka(JP), Angewandte(DE) , U_Utrecth(NL) |
| | polycrystalline-film | | | c7 | d6 | e1 | f5 | AstroPower(US), Sharp(JP) |
| Regular segment | compound-semiconductor | a12 | | c4 | d4 | | | Matsushita(JP), Sony(JP) |
| | amorphous-film | a4 | b1 | | d1 | | | Penn_State(US), Midwest(US), Schott(DE) |
| | deposition | a4 | b1 | | | e4 | | Schott(DE), Hahn-Meitner(DE), Luch(US), Penn_State(US) |
| Minor segment | accumulated-charge | a3 | | | | | f3 | Kaneka(JP) |
| | reflective-film | a7 | | c6 | | | | Nakata(JP) |
| | semiconductor-layer | a11 | | | d5 | | | Canon(JP) |
| | absorber-layer | a6 | | c8 | | | | Kaneka(JP), Hahn-Meitner(DE), Midwest(US), IST-Institut(DE), Showa(JP) |
| | composite-structure | a9 | b6 | | | | | Midwest(US) |
| | light-transmissive | | | | d7 | e3 | | Sharp(JP) |
| Niche segment | encapsulant | a5 | | | | | | Kaneka(JP) |
| | silicon-film | a8 | | | | | | Semiconductor_Energy(JP), Midwest(US) |
| | liquid-coating | | b5 | | | | | McCandless(US) |
| | porous-silicon-layer | | b3 | | | | | Canon(JP) |
| | pn-double-layer | | | c5 | | | | ANTEC_Solar(DE) |
| | anti-reflection-coating | | | | d8 | | | Pacific_Solar(AU) |
| | roll-to-roll-fashion | | | | | e4 | | Luch(US) |
| | thermal-isolation-layer | | | | | e2 | | Industrial_Technology(TW) |
| | CIGSS-absorber | | | | | | f4 | U_Central_Florida(US) |
| | sulfur-compound | | | | | | f2 | Honda_Giken(JP) |
| | CdTe-film | | | | | | f6 | Solar_Systems(IT) |

**(2) Strategic Suggestions:** According to the above table of technology segments, technical topics, and related companies, the relations among three factors could be observed from different viewpoints, including of an industry, technical topics, and companies. The strategic suggestions would be recognized depending on the domain knowledge and described as follows.

Viewpoint of an industry: Referring to Table 1, the most common areas of thin-film solar cell during 2000 till 2009 were H01L031-18, H01L021-02, H01L031-06, H01L031-036, H01L031-00, and H01L021-00, where H01L031 and H01L021 were also indicated as the top two out of the 10 popular areas in the classification search of European Patent Office [19]. It was suggested that most companies should put their efforts in these areas. On the other hand, the less important areas of thin-film solar cell spread widely to 115 categories (each category containing only one patent). It meant that the new directions of this industry emerged rapidly and variedly.

Viewpoint of technical topics: From the above Table 1, the technical topics in the significant segment (i.e., porous-structure, annealing-process, plasma-CVD, pin-layer, and polycrystalline-film) were the most popular and valuable technical items. It was suggested that the government units should pay more attention to these items. The technical topics in the niche segment (i.e., encapsulant, silicon-film, liquid-coating, porous-silicon-layer, pn-double-layer, anti-reflection-coating, roll-to-roll-fashion, thermal-isolation-layer, CIGSS-absorber, sulfur-compound, and CdTe-film) were the emerging technical items. It appeared that some of these items could be potentially new techniques.

Viewpoint of companies: According to Table 1, the companies in the significant segment (e.g., Sony (JP), U-Utrecth (NL), Penn-State (US), Sharp (JP), National-Institute (JP), and Canon (JP) in the "porous-structure" topic) were the more competitive ones. It was suggested that the companies in the same technical topic could cooperate together to form an alliance to increase their strength; but for the strongest company, it might be appropriate to compete with all others to get the leader position. In contrast, the companies in the niche segment were the ones with innovative ideas. It was suggested that these companies should re-examine their R&D plans to decide whether to put more resources into this potential technical topic or to withdraw resources from this unavailable technical topic.

In addition, the focused companies were the ones with higher frequency (equal or greater than 3) in Table 1: Kaneka (JP), Sharp (JP), Midwest (US), Canon (JP), Matsushita (JP), Penn-State (US), and Schott (DE). It would be suitable for these companies to compete for the leader position.

## 5 Conclusions

The research design of IPC classification and association analysis for strategic planning has been formed and applied to propose the strategic suggestions for thin-film solar cell using patent data. The experiment was performed and the experimental results were obtained. The visualized results: association diagrams, crystallized diagrams, and integrated map, were generated. The leading six IPC categories were: H01L031-18, H01L021-02, H01L031-06, H01L031-036, H01L031-00, and H01L021-00. The technical topics, related companies, and technology segments were identified. The technical topics in the significant segment were: porous-structure, annealing-process, plasma-CVD, pin-layer, and polycrystalline-film. The focused companies were: Kaneka (JP), Sharp (JP), Midwest (US), Canon (JP), Matsushita (JP), Penn-State (US), and Schott (DE). Finally, the strategic suggestions on thin-film solar cell were also recognized and proposed.

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 strategic planning 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. Blackman, M.: Provision of Patent Information: A National Patent Office Perspective. World Patent Information 17(2), 115–123 (1995)
2. Tseng, Y., Lin, C., Lin, Y.: Text Mining Techniques for Patent Analysis. Information Processing and Management 43, 1216–1247 (2007)
3. Wikipedia, Strategic planning (October 15, 2010), http://en.wikipedia.org/wiki/Strategic_planning
4. Barney, J.B., Hesterly, W.S.: Strategic Management and Competitive Advantage: Concepts and Cases. Prentice Hall, Englewood Cliffs (2010)
5. Robbins, S., Coulter, M.: Management, 10th edn. Prentice-Hall, Englewood Cliffs (2008)
6. Solarbuzz, Solar Cell Technologies (October 20, 2010), http://www.solarbuzz.com/technologies.htm
7. Wikipedia, Thin film solar cell (October 20, 2010), http://en.wikipedia.org/wiki/Thin_film_solar_cell
8. Jager-Waldau, A.: PV Status Report 2008: Research, Solar Cell Production and Market Implementation of Photovoltaics, JRC Technical Notes (2008)
9. Feldman, R., Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, Cambridge (2007)
10. WIPO, Preface to the International Patent Classification (IPC)(October 30, 2010), http://www.wipo.int/classifications/ipc/en/general/preface.html
11. 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)
12. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison Wesley, Reading (2006)
13. Ohsawa, Y., Benson, N.E., Yachida, M.: KeyGraph: Automatic Indexing by Co-Occurrence Graph Based on Building Construction Metaphor. In: Proceedings of the Advanced Digital Library Conference (IEEE ADL 1998), pp. 12–18 (1998)
14. Maeno, Y., Ohsawa, Y.: Human-Computer Interactive Annealing for Discovering Invisible Dark Events. IEEE Transactions on Industrial Electronics 54(2), 1184–1192 (2007)
15. Maeno, Y., Ohsawa, Y.: Stable Deterministic Crystallization for Discovering Hidden Hubs. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, vol. 2, pp. 1393–1398 (2006)
16. Chiu, T.F.: Applying KeyGraph and Data Crystallization to Technology Monitoring on Solar Cell. Journal of Intelligent & Fuzzy Systems 21(3), 209–219 (2010)
17. USPTO. the United States Patent and Trademark Office (October 30, 2010), http://www.uspto.gov/
18. Stanford Natural Language Processing Group, Stanford Log-linear Part-Of-Speech Tagger (October 15, 2010), http://nlp.stanford.edu/software/tagger.shtml
19. European Patent Office, Search the European classification (October 30, 2010), http://v3.espacenet.com/eclasrch?classification=ecla&locale=en_EP