A SAO-based Approach for Technology Evolution Analysis Using Patent Information: A Case Study on Graphene Sensor

Zhengyin Hu[1,2] and Shu Fang[1]

huzy@clas.ac.cn
[1]Chengdu Documentation and Information Center, Chinese Academy of Sciences,
No.16, Nan'erduan, Yihuan Road, Chengdu (China)
[2]University of Chinese Academy of Sciences,
No.19A Yuquan Road, Beijing (China)

[1]fangsh@clas.ac.cn
[1]Chengdu Documentation and Information Center, Chinese Academy of Sciences,
No.16, Nan'erduan, Yihuan Road, Chengdu (China)

Introduction

The Subject-Action-Object (SAO) structures are composed of Subject (noun phrase), Action (verb phrase) and Object (noun phrase), which can represent technology information with more details in a simple manner and have been widely applied in patent text mining (Cascini, Lueehesi, & Rissone, 2001; Sungchul et al., 2012; Zhang et al., 2014a). This paper presents an approach for technology evolution analysis based on SAO. SAO structures are extracted and cleaned from patent text. The technology information of patents such as problems, solutions, functions and effects are stated by SAO. By calculating the distributions of problems over solution groups, a technology evolution map of problems can be drawn. Graphene sensor patents are selected as a case study.

Methodology

Extracting SAO Structures

After collecting patents, some national language processing (NLP) tools are used to extract raw SAO structures from patent text fields. Normally, the fields such as "Title" and "Abstract" are precise and meaningful for NLP (Sungchul et al., 2012).

Cleaning SAO Structures

The number of raw SAO structures is huge and they need to be cleaned. Text mining tools and domain thesauri are used to carry out Subject and Object cleaning by following a term clumping framework (Zhang, et al., 2014b). The verb phrases of Action are normalized and categorized by experts.

Tagging SAO Structures

According to a classification model learned from a training data, the cleaned SAO structures are tagged with 4 kinds of labels of *problem*, *solution*, *function* and *effect*.

Clustering SAO of Solution

After tagging the semantic type of each SAO, those with *solution* label are clustered into different *solution* groups. Each solution group with similar SAO can be considered as a *solution* topic.

Drawing technology evolution map of problems

Kim, Suh and Park (2008) approached a method that can be used to draw technology evolution map of keywords by calculating the distributions of keywords over the keyword cluster groups. We draw technology evolution map of problems based on Kim, Suh and Park's (2008) research. Firstly, we calculate the distributions of problems over the *solution* groups. If the co-occurrence frequency of two problems is above a threshold, we draw a directed line segment between them to show their relevance. Then the occurrence frequency of each problem in *solution* groups is counted. Finally, by adding the earliest filling date of each problem, a technology evolution map of problems with horizontal axis of timeline and vertical axis of frequency can be drawn.

Case Study

Extracting SAO Structures

We selected Derwent Innovations Index (DII) as data source and invited experts to determine the patent retrieval strategy for graphene sensor patents. After eliminating irrelevant patents, we got 196 patents. We extracted raw SAO from the "Title" and "Abstract" fields and got 4,823 raw SAO structures using an NLP tool named ReVerb (Anthony, Stephen & Oren, 2011).

Cleaning SAO Structures

We cleaned Subject and Object by using a commercial text mining tool, VantagePoint (Nils, 2011) and domain thesauri. We followed the term clumping framework to clean them, which includes general cleaning, terms pruning and terms consolidating processes. After term clumping, we got 628 terms of Subject and Object. We normalized and categorized the verb phrases of Action based on a rule table made by experts. After the cleaning steps, we got 2250 SAO structures.

Tagging SAO Structures

We chose 167 SAO structures from 20 patents as a training set. We picked up Subject, Action as the classification features and C4.5 decision tree as the classifying algorithm to build a classification model which helps to categorize SAO to 4 classes of *problem*, *solution*, *function* and *effect*. Among the classified SAO structures, there are 208 tagged with *problem* label, 746 with *solution* label, 824 with *function* label and 472 with *effect* label. A sample of SAO is shown in table 1.

Clustering SAO of Solution

We clustered the SAO structures with *solution* label into *solution* groups using *k*-means algorithm. By comparing the cluster results, we set the *k*-value 20 and got 20 *solution* groups.

Drawing technology evolution map of problems

By calculating the distributions of problems over each *solution* group, a technology evolution map of problems in graphene sensor patents was drawn. A part of the map is shown in Figure 1.

Table 1. A sample of SAO after tagging.

Type	Subject	Action	Object
Problem	method	synthetize	graphene oxide
Solution	method	use	ultrasonic oscillation process
Solution	graphite powder	mixed with	sodium nitrate
Function	graphene oxide	used for	thin film transistor

Conclusions

The technologies in the upper left corner of Figure 1 appeared in many different *solution* groups and were applied for patents in earlier time, which can be considered as the basic problems in graphene sensor, such as ***producing carbon nanotube***, ***synthetizing graphene oxide***, etc. The technologies in the lower right corner of Figure 1 appeared in fewer *solution* groups and were applied for patents lately, which can be considered as the latest technologies or emerging technologies, such as ***manufacturing sensor array***, ***detecting nucleic acid***, etc**.**

We can draw a technology evolution map of solution, function or effect by following a similar process. The separate technology evolution maps of problem, solution, function and effect can be combined to a more comprehensive technology evolution map of graphene sensor. This study is ongoing.

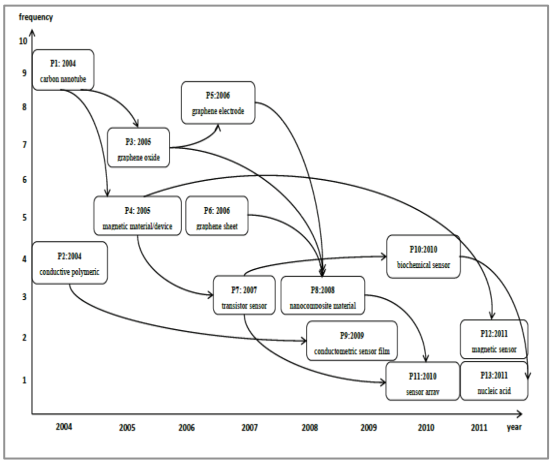

Figure 1. A part of technologies evolution map of problems in graphene sensor patents.

References

Anthony, F., Stephen, S., & Oren E. (2011). *Identifying Relations for Open Information Extraction*. Retrieved March 2, 2014 from: http://ai.cs.washington.edu/www/media/papers/reverb.pdf.

Cascini, G., Lucchhesi, D. & Rissone, P. (2001). Automatic patents functional analysis through semantic processing. *The 12th ADM International Conference*. Rimini, Italy.

Nils N. (2011). *VantagePoint*. Retrieved April 24, 2015 from: https://www.thevantagepoint.com/data/documents/VP%20INTRO%202011.pdf.

Sungchul, C., Hyunseok, P., Dongwoo, K., Lee, J.Y., & Kim, K. (2012). An SAO-based text mining approach to building a technology tree for technology planning. *Expert Systems with Application*, *39*, 11443-11455.

Kim, Y.G., Suh, J.H. & Park, S.C. (2008). Visualization of patent analysis for emerging technology. *Expert Systems with Applications*, *34*, 1804–1812.

Zhang, Y., Porter, A. L., Hu, Z., Guo, N., & Newman, N.C. (2014b). “Term clumping” for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change*, *85*, 26-39.

Zhang, Y., Zhou, X., Porter, A. L., & Gomila, J. (2014a). How to combine term clumping and technology roadmapping for newly emerging science & technology competitive intelligence: The semantic TRIZ tool and case study. *Scientometrics, 101*(2), 1375-1389.

Prediction of Potential Market Value Using Patent Citation Index

HeeChel Kim[1,2], Hong-Woo Chun[2], Byoung-Youl Coh[2]

{kim, hw.chun, cohby}@kisti.re.kr
[1]University of Science and Technology, 305-350, 217 Gajeong-ro, Yuseong-gu, Deajeon(South Korea)
[2]Korea Institute of Science and Technology Information, Dept. Of Technology Intelligence Research, 130-741,
66 Hoegiro, Dongdaemun-gu, Seoul (South Korea)

Introduction

Patent statistics have frequently been used as both technological and economic indicators, however, in order to fully utilize patent data in economic analyses, we must link patents to economic activity at a level of industry or product.

Many previous pieces of research showed the effectiveness of patents citation index (PCI), containing annual citation information, on economic indicators of respective firms. Hall et al. (2005) have studied the relation between a market value and PCI using the Tobin's q approach, and Patel and Ward (2011) have compared the stock market value of firms with the patent citation using the event study methodologies. Both studies showed that Patent statistics can be effectively used to micro-level economic analyses and the increase of PCI has the positive effect on the corresponding market value.

Meanwhile, our study aims to prove the effectiveness of PCI on the economic value of industry, so-called Meso-level study and, in this case, it is essential to develop technology-industry concordance method.

Method

The correlation analysis between Potential Market value (PMV) and PCI for the respective industry is carried out in three stages.

(1) Data concordance process. The market data was collected from Annual Survey of Manufactures (ASM) [1] in the US Census Bureau (http://www.census.gov) and PCI [2] data was collected from the patent set registered USPTO.

Next, we created an annual concordance matrix of IPC (international patent classification) 4-digit to NAICS (North American industry classification system) 6-digit (rev.2002, 2007, and 2012) by Algorithmic Links with Probabilities (ALP), ALP (Lybbert & Zolas, 2013), concordance method of the WIPO (http://www.wipo.int/). ALP is the most up-to-date method compared with those of YTC (Kortum & Putnam, 1997), OECD (Johnson, 2002) and DG (Schmoch et al., 2003).

Each IPC 4-digit is connected to multiple NAICS 6-digit probabilistically via a text mining-based matching rule.

PMV was calculated by model 1 as follows, and consequently, 593 annual pairs of PMV-PCI for each IPC were generated.

$$PMV_{ij} = \frac{\sum_{k=1}^{476} a_{ijk} \times b_{ik}}{\sum_{i=1}^{593} \sum_{k=1}^{476} a_{ijk} \times b_{ik}} \times \sum_{k=1}^{476} b_{ik} \text{ …… Model 1.}$$

a = Probability of IPC 4-digit to NAICS 6-digit
b = Value of shipment by NAICS in ASM
i = Year (2002 to 2013)
j = IPC 4-digit code (A01G, A01H, …, H05K)
k = NAICS 6-digit code (311111, 311119, …, 339999)

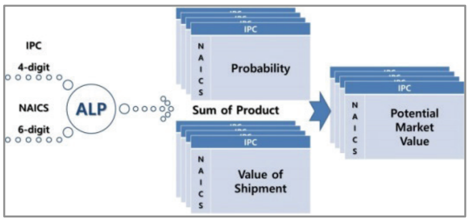

Figure 1. Process of IPC-NAICS Concordance and PMV Calculation.

(2) Statistical correlation analyses for all industry fields. We performed a statistical correlation analysis between the annual incremental of PMV and PCI. We used the Spearman's rho correlation analysis, a nonparametric correlation analysis algorithm, useful to calculate the correlation between the ranked variables (IBM, http://k:5172/help/index.jsp?topic=/com.ibm.spss.statistics.tut/introtut2.htm).

(3) Statistical correlation analyses for 4 major industry fields. The correlation analyses between the annual incremental of PMV and PCI for 4 major industry fields - electrical engineering, instruments, chemistry, and mechanical engineering – were also performed.

Result

Figure 2 shows annual trends of PMV, PCI, and Patent registered. All kinds of variables are trending upward in an accelerating degree.

[1] ASM is estimated sample statistics issued annually for more than one people employees firms in the manufacturing sector. ASM is classified industries by NAICS. In this study, using field of the value of shipment at the 2004 and 2006 edition of ASM that follow the revised NAICS 04 and 2008 to 2011 edition of ASM that follow the revised NAICS 07.

[2] PCI data was used granted patent of USPTO. During the year of from 2002 to 2013.

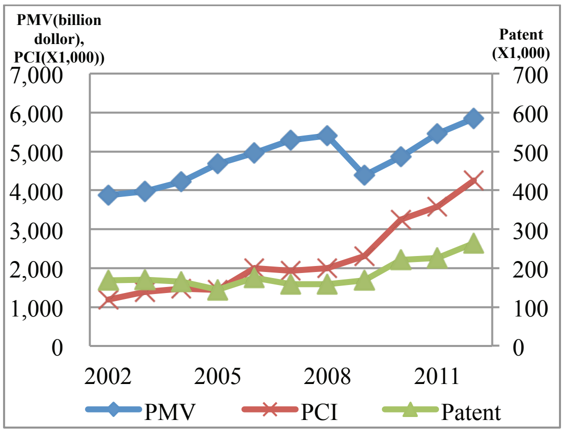

Figure 2. Structure of PMV, PCI and Patent.

PMV of each IPC

Table 1 shows the result of the PMV of each IPC calculated from model 1. It has a significant meaning that a set of patents can be expressed to market value.

Table 1. PMV (unit: million US$).

No.	IPC	2002	2003	…	2013
1	A01G	282	301	…	229
2	A01H	3,057	3,831	…	15,227
...	…	…	…	…	…
593	H05K	6,556	6,166	…	5,055

Correlation Analyses

In the analysis results over the entire industry fields (Table 2), we could find out that significance of correlation and direction varies depending on the Lagging time (differences in data collection year between PMV and PCI). It has a relatively weak positive correlation when the lagging time is 0, meanwhile, it showed relatively strong negative correlation when the lagging time is "PCI+1" – the data collection year for PCI is one year after to that of PMV - . And in case of the lagging time of "PCI-1", it has relatively strong positive correlation, which reveals patent citation activity's positive relation to the corresponding market value "one year later".

Table 2. Results of PMV-PCI rate's correlation analyses (all fields, **significance level 0.01).

Lagging time(year)	Correlation coefficient	p-value (two-tailed)	N
PCI-1	0.136**	0.000	5337
0	0.093**	0.000	5930
PCI+1	-0.323**	0.000	5337

The analyses results of 4 major industry fields showed similar tendencies to all-field-analysis except electrical engineering field.

Table 3. Results of PMV-PCI rate's correlation analyses (4 major fields, **significance level 0.01).

Field	Lagging time(year)	Correlation coefficient	p-value (two-tailed)
Electronic	PCI-1	-0.013	0.747
	0	0.143**	0.000
	PCI+1	-0.513**	0.000
Instrument	PCI-1	0.209**	0.000
	0	0.011	0.795
	PCI+1	-0.360**	0.000
Chemistry	PCI-1	0.180**	0.000
	0	0.022	0.434
	PCI+1	-0.265**	0.000
Mechanic	PCI-1	0.167**	0.000
	0	0.123**	0.000
	PCI+1	-0.266**	0.000

Conclusion

In this research, we made a systematic way for describing the technological impact on industry sector by using some indices, which has a significant meaning that a set of patents can be expressed to market value. We also had confirmed the potential of PCI to predict PMV of the industry. Experimental results showed that PMV in all industry fields was related by the corresponding field's patent-citation activity in one year before or after. After this work, we will deal with enhanced concordance approach to find out relationships between IPC 7-digit and NAICS 7-digit. Also, the self-citation ratio of patent-citation activity may affect economic activity at a level of industry or product, which is now on a study.

References

Hall, B. H., et al. (2005). Market value and patent citations. *RAND Journal of Economics*, 16-38.

Johnson, D., March (2002). The OECD Technology Concordance (OTC): Patents by Industry of Manufacture and Sector of Use, OECD Science, Technology and Industry Working Papers.

Kortum, S. & Putnam, J. (1997). Assigning patents to industries: tests of the Yale technology concordance. *Economic Systems Research*, *9*(2), 161-176.

Lybbert, T.J. & Zolas, N.J. (2014). Getting patents and economic data to speak to each other: An 'algorithmic links with probabilities' approach for joint analyses of patenting and economic activity. *Research Policy*, *43*(3), 530-542.

Patel, D. & Ward, M.R. (2011). Using patent citation patterns to infer innovation market competition. *Research Policy*, *40*(6), 886-894.

Schmoch, U., Laville, F., Patel, P., & Frietsch, R. (2003). Linking technology areas to industrial sectors. *Final Report to the European Commission, DG Research*, *1*.