Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/070520

A Predictive Model for Patent Registration Time Using Survival Analysis

Sunghae Jun[1] *and Daiho Uhm*[2,*]

[1] Department of Statistics, Cheongju University, Cheongju, Korea
[2] Department of Statistics, Oklahoma State University, Stillwater, OK, USA

Received: 13 Jan. 2013, Revised: 16 May. 2013, Accepted: 17 May. 2013
Published online: 1 Sep. 2013

Abstract: The infringements and suits of patents have been increased in many technological fields. Since a patent is an intellectual property to protect inventors' exclusive rights of developed technologies, technological pre-occupancy is very important to the development of companies. For the technological pre-occupancy a patent registration time is surely important. In this paper we propose a predictive model of the patent registration time using survival analysis with a Weibull survival regression model and a Cox's proportional hazard model. They are modeled on the number of international patent classifications (IPC) codes and keywords. To verify the proposed models, we perform a case study using the retrieved patent documents of 'hybrid vehicle' from the website of United State Patent and Trademark Office (USPTO).

Keywords: Patent registration, Survival analysis, Patent document data, Predictive model

1 Introduction

Patent is the exclusive right granted by a government to manufacture or sell developed technologies in a given time period. Most companies have tried to apply and register their developed technologies for patents around world. In addition, the infringements and suits of patents between the companies have increased in many technological fields. Therefore it is very important for companies to develop technologies and keep their occupancy as a registered patent. Since a patent registration is a good approach for technological occupancy, it is useful information for a company to know the registered time. Jung et al. [1] suggested a patent registration model using a multi-layer perceptron (MLP) of back propagation. The MLP model could not consider censored data. The censored cases occur in the process of patent registration since all applied patents are not registered and some of them are even rejected. To solve the censoring problem, we propose a predictive model of patent registration time using survival analysis. A survival model is a good approach to deal with the censored data [2]. This research begins with the relationship between patent registration and patent survival time. A Weibull survival regression and a Cox's proportional hazard models are applied to the patent data. We analyze the registration time of patent using the IPC codes and keywords of patent data. To verify the performance of the proposed model, we perform a case study using the retrieved patent documents of 'hybrid vehicle'. Using text mining technique, we preprocess the searched patent data. This paper contributes to R&D planning and technology management of companies and nations.

2 Research Background

2.1 Patent management

Intellectual property management is a process of getting and keeping intellectual properties using legal and business means. Patent management is one of the intellectual property managements since a patent is the intellectual property. The results of developed technology are applied to patents and the companies do their best to register the patents as soon as they can. This is one of the diverse approaches of patent management. In this paper, the patent registration in patent management approaches is studied. Companies should improve competitiveness of their developed technologies since the patent system allows the exclusive right of developed technology to the patent owner.

[*] Corresponding author e-mail: daiho.uhm@okstate.edu, daiho.uhm@hotmail.com

To know the chance and time of patent registration is essential to R&D planning of companies and nations. The aim of this research is to build a predictive model of patent registration time based on survival analysis.

2.2 Survival analysis

Survival analysis takes care of time to an event, which could be a death, equipment breakdown, bankruptcy, or conviction. The times to an event might not be completely observed for some reason. When the times to death by liver cancer are measured, a patient might have died from a car accident, and some patents could be terminated from the study. When the complete times could not be collected, they are called *censored*. Sometime the starting time is not known when patients who already have a disease are surveyed. This incomplete data is called *truncated*. Parametric, semiparametric, and nonparametric schemes are used to estimate survival functions and hazard rates. The Kaplan-Meier estimator [3] for the survival function and the Nelson-Aalen estimator [4] [5] for the cumulative hazard function are widely used as non-parametric estimators. Let $F(\cdot)$ and $f(\cdot)$ be a cumulative distribution and probability density function for the time to an event, respectively. The survival function is defined by

$$S(t) = 1 - F(t) = P[T > t] = \int_t^\infty f(x)dx \,. \tag{1}$$

It is the survival probability after time t. The hazard rate is defined as the probability of risk just after time t,

$$h(t) = \lim_{\triangle t \to 0} \frac{P[t \leq T < t + \triangle t | T \geq t]}{\triangle t} \tag{2}$$

and the cumulative hazard function is defined by

$$H(t) = \int_0^t h(x)dx \,. \tag{3}$$

The survival function and the cumulative hazard function have the following relationship;

$$S(t) = exp\{-H(t)\} \,. \tag{4}$$

For parametric estimators an exponential, weibull [6] [7], log normal, or gamma distribution could be used depending on the nature of parameters and the shape of the survival function.

3 Proposed predictive model for patent registration time

Survival models are proposed for predicting the time of patent registration. It is important to know the registration time for the management of technology (MOT). In MOT, patent is one of the final products of R&D works. Technological competitiveness depends on the quantity and quality of patents among companies and countries. Most companies would like to register their patents first since patent system protects only the first inventors' right. Therefore companies are interested in predicting the registration time for their MOT.

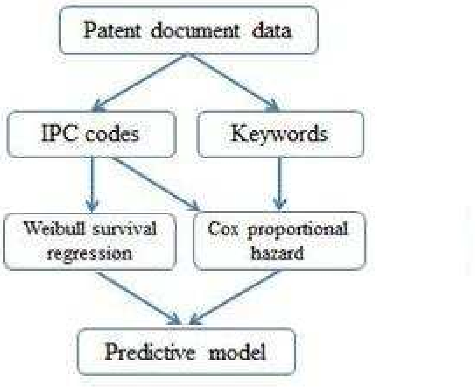

Fig. 1 Process of proposed model

In this paper, we construct a predictive model using survival analysis to study patent registration time. We consider a Weibull survival regression and Cox proportional hazard models which are parametric and semiparametric survival regression models, respectively. First, we search for patent documents of a given technology field. Second, the retrieved patent data are preprocessed by text mining techniques. We calculate the numbers of IPC codes and keywords from the preprocessed data set. Third, the two survival approaches are applied to the data set. The Weibull survival regression and Cox's proportional hazard models are applied to the IPC data set, and the keywords data is analyzed by Cox's model. Figure 1 shows the diagram of the processes for the proposed model. Finally, we combine the results for constructing our predictive model. The dependent (response) variable in the survival models is defined as the time difference between applied and registered dates in days; the number of IPC codes and the frequency of keywords in each patent document are the explanatory (independent) variables.

3.1 Weibull survival regression model

Let $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_p)^t$ be a vector of explanatory variables. Covariate $Z_1 = 1$ might be allowed as an intercept. The time to event, T, could be modeled by

$$ln(T) = \beta^t \mathbf{Z} + \sigma W \,,$$

where ln is a natural logarithm function, β is a vector of coefficients, and W has Weibull distribution with parameters $\lambda > 0$ (a scale parameter) and $\alpha > 0$ (a shape parameter). The survival function of the failure time T is given by

$$S_T(t|\mathbf{Z}) = exp\left[-\left\{t \cdot exp(-\beta^t \mathbf{Z})\right\}^\alpha\right] \,.$$

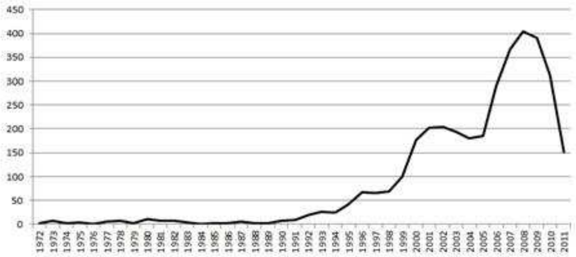

Fig. 2 Number of patents regarding hybrid car by year

The maximum likelihood estimators (MLE) for the coefficients, β, are provided in many statistical packages.

3.2 Cox proportional hazard model

Cox [8] proposed a hazard rate depending on covariates. Let $h(t|\mathbf{Z})$ be the hazard rate for an individual with covariates, $\mathbf{Z}$, at time t. Then the Cox's proportional hazard model is

$$h(t|\mathbf{Z}) = h_0(t)exp(\beta^t\mathbf{Z}) \,, \tag{5}$$

where $h_0(t)$ is a baseline hazard. Based on the hazard rate in (5) a partial likelihood is provided which the MLEs are estimated by.

The baseline survival function, $S_0(t)$, is estimated by the equations (3) and (4);

$$\hat{S}_0(t) = exp\{-\hat{H}_0(t)\} \,.$$

It is an estimated survival function with $\mathbf{Z} = \mathbf{0}$. The survival function of an individual with $\mathbf{Z} = \mathbf{Z}_0$ is estimated by

$$\hat{S}(t|\mathbf{Z}_0) = \hat{S}_0(t)^{exp(\beta^t\mathbf{Z}_0)} \,.$$

4 Experiment and Result

In this experiment, patent documents are retrieved from the website of the United State Patent and Trademark Office (USPTO, www.uspto.gov) using the following keywords equation;

Abstract = hybrid*(car + vehicle + automobile) .

Figure 2 shows the number of the patents by year. It is known that hybrid car technology was not developed actively until early in the 1990s. However the number of patents increased rapidly between 1998 and 2008. Currently, the technology is still developed by lots of motor vehicle companies.

Until June 27, 2012, there are 3,574 patent documents regarding *hybrid car*, which have abstracts including 'hybrid car', 'hybrid vehicle', or 'hybrid automobile'. There are 1,468 applied patents which are not registered yet, so they are all right-censored in the survival analysis.

Table 1 Frequency table of number of IPC codes

nipc	Freq.	Percent	Censored
1	2204	61.67	750
2	772	21.60	380
3	351	9.82	212
4	137	3.83	79
5	61	1.71	25
6	22	0.62	11
7	13	0.36	6
8	7	0.20	3
9	3	0.08	1
10	2	0.06	0
11	2	0.06	1
Total	3,574	100.00	1,468

Table 2 Fitted model by Weibull survival regression on number of IPC codes

Parameter	MLE	S.E.	Chi-square	p-value
Intercept	7.4214	0.0253	86,040.5	$< .0001$
nipc	0.0500	0.0137	13.33	0.0003
AIC = 7,059.807				

For data analysis and statistical computing, SAS 9.3 (www.sas.com) and R language (www.r-project.org) are used. Also the *tm* package in R is applied in the text mining of the patent documents.

4.1 Using the number of IPC codes

The registration times (in days) of the patents in the hybrid car technology are analyzed by Weibull survival regression and Cox's proportional hazard models. In Table 1, the frequencies and percents of the number of IPC codes (*nipc*). Also, the number of censored observations are given for the 3,574 patent documents.

In Table 2, the registration times are fitted on the *nipc* in each patent by the Weibull survival model. The covariate of *nipc* has a strongly significant (p-value = 0.0003) effect to the patent registration time with MLE 0.0500. It would be interpreted that the registration time is increased with an increased number of IPC codes.

However, there are more than 60% of patent documents with only one IPC code in Table 1. The Weibull survival model and Cox's proportional hazard model are fitted based on a two-grouped variable which is defined with only one IPC code (*grp* = 0) and with two or more codes (*grp* = 1).

In Table 3, the estimated parameter for *grp* by the Weibull survival model are strongly significant (p-value $< .0001$) with MLE 0.1285. It is interpreted that the registration time for a patent is longer with two or more IPC codes than with only one code. There is another parameter estimate (MLE) of the coefficient from Cox's proportional

Table 3 Fitted models by two-grouped variable

Parameter	MLE	S.E.	Chi-square	p-value
	Weibull survival regression model			
Intercept	7.4586	0.0168	196436	$< .0001$
grp	0.1285	0.0305	17.70	$< .0001$
	AIC = 7,056.034			
	Cox's proportinal hazard model			
grp	-0.2941	0.0472	38.83	$< .0001$
	AIC = 31,389.693			

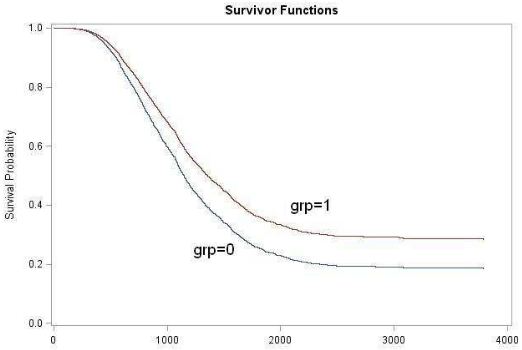

Fig. 3 Survival function of only one IPC code (*grp*=0) and two or more codes (*grp*=1) by Cox's hazard model (in days)

hazard model which is -0.2941 with p-value $< .0001$, and in Figure 3 survival curves are plotted by the model. The hazard rate which was defined in (2) could be interpreted as a probability that a patent is registered immediately at time t, and the survival function in (1) suggests the probability that the registration takes longer than a given time t. With the negative coefficient of -0.2941 the hazard rates are estimated by

$$\begin{aligned} h(t|Z=0) &= h_0(t) \text{ ,and} \\ h(t|Z=1) &= h_0(t)exp(-0.2941) \\ &= 0.7452h_0(t) \text{ .} \end{aligned}$$

It could be explained that an applied patent with only one IPC code ($Z=0$) has a higher probability to be registered immediately at time t than with two or more codes ($Z=1$). In the Figure 3, the survival probability of the patents with only one IPC code (*grp*=0) is smaller than that with two or more codes (*grp*=1) at a given time t. Cox's hazard model in (5) and the survival functions suggest that it takes longer time to register a patent when the application has two or more IPC codes.

When the Akaike Information Criterions (AIC) are compared among the three models, the Weibull survival model with the two-grouped variable has the best (smaller is better) AIC of 7,056.034.

Finally, we conclude that registration time is increased with an increased number of IPC codes.

Table 4 Extracted keywords of the hybrid car patent documents

Keyword	Mean frequency	Max. frequency
air	0.17	10
battery	0.65	22
charge	0.19	12
clutch	0.29	11
combustion	0.52	12
electric	1.46	19
energy	0.56	15
fuel	0.37	15
gear	0.30	17
generator	0.29	10
internal	0.43	12
pressure	0.14	11
shaft	0.29	19
speed	0.38	12
storage	0.30	10
temperature	0.14	10
torque	0.60	16
transmission	0.46	12
voltage	0.22	14

Table 5 The selected three keywords by forward selection using Cox's hazard model

parameter	MLE	S.E.	Chi-square	p-value
electric	0.0238	0.0104	5.2518	0.0219
energy	0.0273	0.0125	4.7205	0.0298
shaft	-0.0764	0.0246	9.6817	0.0019

4.2 *Using the frequency of keywords*

In this section, Cox's proportional hazard model is fitted with technological keywords of the hybrid car patent documents. The top ranked and meaningful keywords are extracted from the patent data set. Table 4 shows the extracted keywords with their mean frequencies in a patent document, and the maximum frequencies of each keyword. The keyword *electric* is most frequent with an average of 1.46 times, and *battery* was used 22 times in a single document.

The most significant three keywords are selected by forward selection using Cox's proportional hazard model. In Table 5 the keywords of *electric*, *energy*, and *shaft* are selected at a 5% significance level. Positive estimates for *electric* and *energy* mean that increased use of those keywords in a patent document increases the probability to register immediately. A negative coefficient means using the keyword *shaft* in a patent document decreases probability to register immediately.

5 Conclusion

The registration time is very important since earlier application and registration of patents provides companies better competitiveness. Therefore many companies want to predict and shorter their patent registration time.

We proposed predictive models to analyze the time of patent registration. The Weibull survival regression model and Cox's proportional hazard model are applied to analyze the data set for patent registration time. We carried out a case study to show how survival models could be applied to real problems using the retrieved patent documents related to *hybrid car*. The survival models considered the number of IPC codes and the number of extracted keywords in each patent document as explanatory variables. The dependent variable was defined as the time difference between applied and registration date in days.

We found that the patents with only one IPC code were registered earlier than these with two or more codes, and that the frequency of some keywords significantly affected the patent registration time. In future works, more survival models would be studied for predicting patent registration time and forecasting technology trends.

Acknowledgement

The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] W. Jung, S. Park, and D. Jang, Patent Registration Prediction Methodology Using Multivariate Statistics, IEICE Transactions on Information and Systems E94-D-11, 2219-2226 (2011).

[2] J. P Klein, and M. L Moeschberger, Survival Analysis: Techniques for Censored and Truncated Data, Springer, New York, (2010).

[3] E. L Kaplan, and P. Meier, Nonparametric Estimation from Incomplete Observations, Journal of the American Statistical Association, **53**, 457-481 (1958).

[4] W. Nelson, Theory and Application of Hazard Plooting for Censored Failure Data, Technometrics, **14**, 945-965 (1972).

[5] O. O Aalen, Nonparametric Inference for a Family of Counting Process, Annals of Statistics, **6**, 701-726 (1978).

[6] W. A Weibull, Statistical Theory of the Strength of Materials, Ingeniors Vetenskaps Akakemien Handlingar, **151**, 293-297 (1939).

[7] W. A Weibull, A Statistical Distribution of Wide Applicability, Journal of Applied Mechanics, **18**, 293-297 (1951).

[8] D. R Cox, Regression Models and Life Tables (with Discussion), Journal of the Royal Statistical Society B, **34**, 187-220 (1972).

[9] T. Feinerer, K. Hornik, and D. Meyer, Text mining infrastructure in R, Journal of Statistical Software, **25**, 1-54 (2008).

[10] R Development Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. www.r-project.org, (2011).

[11] SAS, Strategic Application Software, www.sas.com, (2012).

Sunghae Jun received the BS, MS, and PhD degrees in department of Statistics, Inha University, Korea, in 1993, 1996, and 2001. Also, He received PhD degree in department of Computer Science, Sogang University in 2007. He is currently Assistant Professor in department of Statistics, Cheongju University, Korea. He has researched statistical learning theory and evolutionary algorithms.

Daiho Uhm received his Ph.D. in department of statistics, Florida State University, U.S.A. in 2007, and BS and MS degrees in department of statistics, Inha University in 1997 and 1999, respectively. Currently he is a visiting assistant professor in department of statistics, Oklahoma State University. He is interested in survival analysis and computational statistics.