A Forecasting Model for Technological Trend Using Unsupervised Learning

Sunghae Jun

Depatment of Bioinformatics and Statistics, Cheongju University,
360764 Chungbuk, Korea
shjun@cju.ac.kr

Abstract. Many results of the developed technologies have applied for patents. Also, an issued patent has exclusive rights granted by a government. So, all companies in the world have competed with one another for their intellectual property rights using patent application. Technology forecasting is one of many approaches for improving the technological competitiveness. In this paper, we propose a forecasting model for technological trend using unsupervised learning. In this paper, we use association rule mining and self organizing map as unsupervised learning methods. To verify our improved performance, we make experiments using patent documents. Especially, we focus on image and video technology as the technology field.

Keywords: Image and video technology, Technology forecasting, Unsupervised learning, Association rule mining, Self organizing map.

1 Introduction

In this paper, we propose a forecasting model for technological trend using unsupervised learning. Association rule mining (ARM) and self organizing map (SOM) are popular methods of unsupervised learning. These are used for constructing our forecasting model. Also, we focus on image and video technology (IVT) as a given technological field for our case study. IVT of multimedia database systems has been used an important tool for providing information to humans [1-2]. Many researches of coding, processing, visualization, analysis, and retrieval have been developed in IVT. Recently, biometrics and forensics are needed the IVT [3]. In general, the results of researched and developed technologies for IVT have been published as patent and paper. These are massive literatures. It is difficult to construct forecasting model by them using the quantitative methods of statistics and machine learning [4]. Most technology forecasting (TF) models have been depended on the qualitative methods such as Delphi [5-8]. These qualitative TF are not stable because they are based on subjective knowledge of domain experts. So, we need more objective TF method for IVT forecasting. In this paper, we use a combine model by ARM and SOM to construct a quantitative method for IVT forecasting. ARM is a popular predictive method based on conditional probability [9-10]. SOM is a typical clustering model in unsupervised neural networks of machine learning [11-12]. ARM

T.-h. Kim et al. (Eds.): DTA/BSBT 2011, CCIS 258, pp. 51–60, 2011.

and SOM will create a synergy effect each other to forecast IVT effectively. We will use patent document about IVT until now for IVT forecasting. By analyzing the IVT patent documents, we forecast especially vacant areas of IVT using patent data of U.S, Europe, and China from USPTO (United State Patent and Trademark Office, www.uspto.gov). In next section, we review the related works of IVT and TF. We introduce proposed method to forecast vacant technology of IVT in section 3. To verify the performance our work, we will show experimental results in section 4. Final section includes our conclusions and future works.

2 Image and Video Technology Forecasting

Image technology includes the techniques of material application and management method to create, storage, and analyze images. Based on the image technology, video technology is to capture, record, manage, analyze and reconstruct the continuous images of motion [1-3]. So, IVT is a combined technology based on image and video techniques. TF is to predict a moving trend of technological change. We have R&D plan to avoid infringement of intellectual property using TF results. TF also support meaning knowledge for technology marketing and reducing risk of R&D investment in company and government [13-15]. In this paper, we find vacant TF to give company and government the technological feasibility of each vacant aspect in IVT. The important goal of TF is to monitor the technological trend of given technology [15]. Most TF works have used qualitative and subjective methods depended on experts' prior knowledge such as Delphi [5-8]. TF results of the method are not stable. So, we need more quantitative and objective methods than previous approaches [14]. Some approaches such as roadmap and bibliometrics, have been introduced for TF [14],[16-17]. Typical data source for constructing TF model are paper and patent databases. We use applied patent documents of IVT to forecast vacant technology of IVT in this paper. Principal component analysis (PCA) and SOM were the methods studied on previous TF researches [14],[18-19]. In this paper, we propose a combined TF model including ARM and previous TF method by SOM. We will improve the performance of vacant TF using the proposed methods.

3 IVT Forecasting Using ARM and SOM

Retrieved patent data are so large. We need an analytical method for large patent documents. ARM is a popular method for discovering knowledge from large databases. ARM mines frequent item sets to find meaningful relationship between variables. ARM analyzes item and transaction data sets [9]. $I=\{i_1, i_2, \ldots, i_n\}$ is a set of n binary items. $T=\{t_1, t_2, \ldots, t_m\}$ is a set of transactions. A transaction of T consists of unique identical number and items. A rule of ARM is represented as follows.

$$X \rightarrow Y \quad (X \cap Y = \phi) \tag{1}$$

Where, X and Y are in I. Also X and Y are antecedent and consequent of the rule. We can get so many rules from ARM results. To select meaningful rules from all possible rules, we need criteria to evaluate all rules. Support and confidence are well known constraints. We select the rules satisfied minimum threshold on support and confidence. Support is defined as follow.

$$\text{support}\ (X \rightarrow Y) = \frac{nmber\ of\ transactions\ containing\ both\ X\ and\ Y}{total\ number\ of\ transactions} \tag{2}$$

So, support$(X \rightarrow Y)$ is equal to probability $P(X \rightarrow Y)$. confidence$(X \rightarrow Y)$ is represented by probability $P(Y|X)$ as follow.

$$\text{confidence}(X \rightarrow Y) = P(Y \mid X) = \frac{P(X \cap Y)}{P(X)} = \frac{\text{support}(X \rightarrow Y)}{\text{support}(X)} \tag{3}$$

Lift is additional interest measure in ARM. We can filter the rules satisfied minimum support and confidence constraints to select more meaningful rules [14]. This is defined as follow.

$$\text{lift}(X \rightarrow Y) = \frac{\text{support}(X \Rightarrow Y)}{\text{support}(X)\text{support}(Y)} \tag{4}$$

We get more significant rules by lift measure. Next, the relationship between items is shown by lift value.

$$\text{lift}(X \rightarrow Y)\, value = \begin{cases} > 1, & X\ and\ Y\ are\ complementary \\ 1, & X\ and\ Y\ are\ independent \\ < 1, & X\ and\ Y\ are\ substitutive \end{cases} \tag{5}$$

When the lift value equals to 1, item X and Y are independent each other. That is, X and Y are not associated. The relationship between X and Y is more complementary according as the value is increased over 1. Also, X and Y are substitutive if the value is under 1. In this paper, we use support, confidence, and lift as measures for finding meaningful rules to forecast the IVT. Retrieved patent documents and extracted terms are used as transactions and items of ARM. From the ARM results, we get the relationships between detailed technologies of IVT. Self organizing map (SOM) is a competitive neural networks for classification and clustering [19]. SOM consists of two layers, input and feature. We cluster all patent documents to 2×2 feature map. First, SOM normalize input vector x and initialize weight matrix M. Second, we compute the distance between input and weight using Euclidean distance. The closest m_j to x_i is updated as follow [20].

$$m_k = m_j + \alpha(x_i - m_j) \tag{6}$$

Where, m_j and m_k are current and new weights. So m_k moves to x_i. The constant α is a learning rate to control the speed of converging optimal point. SOM clustering is repeated until satisfying given conditions. We cluster all patent documents to feature map. We find the vacant technology in the feature map with assigned patent documents. In the feature map of SOM, the areas with low density are defined as vacant technology.

In this paper, we use ARM and SOM for forecasting technology of IVT. Our proposed TF approach consists of four steps. First, we prepare data set for constructing IVT forecasting model. We prepare IVT patent documents as analyzed data by ARM and SOM. We retrieve IVT patent documents from USPTO. There is a problem in first step. In general, patent documents are not suitable to analytical methods of statistics and machine learning such as ARM and SOM. To solve this problem, we preprocess the documents using text mining techniques. The output of first step is a document-term matrix (DTM) preprocessed. Our DTM is a structured data to be analyzed. Second, ARM is used for IVT forecasting. We extracted top ranked keywords from DTM, output of first step. Patent document and keyword are transaction and item for ARM. We construct a set of ARM rules by given support and confidence values. In this paper, we determine the values as small as possible because we wish to get ARM rules as many as possible. We search meaningful rules from the set of ARM rules and determine three rules with the highest lift, confidence, and support values. These rules are used to forecast IVT. Third, we use SOM to cluster IVT patent documents. The documents are assigned to the feature map of SOM. In general, the number of clusters is proportional to the dimension of feature map. We perform SOM clustering from large dimension of the feature map to small dimension. In the last SOM result, we find a cluster with vacant technology. The cluster is relatively small but not sparse. Top ten terms form the patent document of the cluster define as vacant area are extracted for defining the vacant technology for IVT. Fourth, we forecast the technology for IVT using the outputs of second and third steps. The following process shows our proposed model for IVT forecasting step-by-step.

```
Technology Forecasting for Image and Video Technology
 Step1. Preparing data set for forecasting IVT
   (1-1) Determining keywords equation for retrieving;
   (1-2) Retrieving IVT patent documents from USPTO;
   (1-3) Preprocessing documents using text mining;
   (1-4) Getting document-term matrix from (1-3)
result;
   (1-5) Dividing DTM into training and test data sets;
   (Output) DTM divided into training and test data.
 Step2. Extracting ARM rules
   (2-1) Extracted top ranked keywords from output of
           step1;
   (2-2) Constructing ARM rules;
   (2-3) Searching meaningful rules by ARM criteria;
      Rule1 by the heighted lift value,
      Rule2 by the heighted confidence value,
      Rule3 by the heighted support value.
   (Output) Three ARM rules - Rule1,2,3.
 Step3. Clustering DTM
   (3-1) Performing SOM clustering from large dimension
          of feature map to small dimension of feature
          map;
   (3-2) Determining optimal number of clusters;
   (3-3) Finding a cluster with vacant technology;
   (3-4) Extracting top ten terms from the patent
          documents included to vacant area;
   (Output) Ten terms extracted from patent documents
in
              the cluster as vacant technology area.
 Step4. Forecasting technology related to IVT using
         outputs from step2 and step3;
```

Also, next figure simplifies the proposed process. A rectangle is an output of each step and an arrow represents the analytical work between outputs.

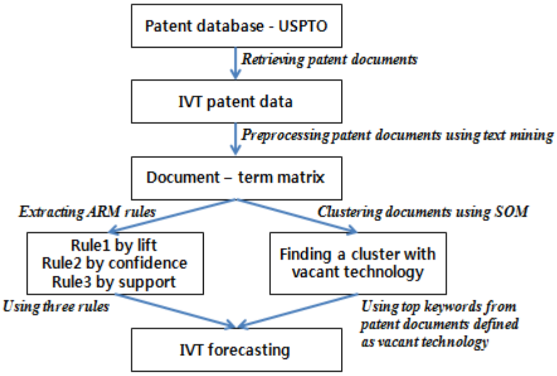

Patent database - USPTO
Retrieving patent documents
IVT patent data
Preprocessing patent documents using text mining
Document – term matrix
Extracting ARM rules
Clustering documents using SOM
Rule1 by lift
Rule2 by confidence
Rule3 by support
Finding a cluster with vacant technology
Using three rules
Using top keywords from patent documents defined as vacant technology
IVT forecasting

Fig. 1. Proposed IVT forecasting model

To forecast IVT, we combine ARM and SOM results. ARM gives the relationship between technological terms of IVT. Vacant technology of IVT is defined by SOM result.

4 Experimental Results

To forecast vacant technology of IVT, we retrieved patent data from USPTO using the following keyword equation.

*Title = [(image + video) * (technology + coding + processing + analysis + retrieval + forensic + application)]*

Where '+' and '*' are 'or' and 'and' operators, respectively. At the first, we retrieved 5001 patent documents until June 8, 2011. The data included some patent documents not related to IVT. We got 3780 patent documents after removing the patents not related to IVT. The available percentage of our retrieved patent documents was

75.58%. The first applied patent was shown in 1984. The issued IVT technologies as patent documents were increased in the late 1990s. The rate of increase has been fast recently. But, 2007 through 2010, the number of patents of IVT was decreased. We intend to find this trend of IVT from our experiment. We divided all 3780 patent documents into two data sets as follow.

Table 1. Dividing all available patent documents into training and test data sets

Data set	Year	Number of patents
Training	1984 – 2005	2251
Test	2006 – 2010	1529
Total	1984 – 2010	3780

Table 2. Occurred frequency levels of top 30 terms

Terms	Occurred frequency levels		
	Low (0 – 1)	Middle (2 – 3)	high (4 –)
allocation	2249	1	1
audio	2227	5	19
beam	2247	2	2
binary	2215	23	13
calibration	2245	4	2
chromaticity	2237	2	12
conversion	2141	66	44
diffusion	2237	11	3
digital	2174	50	27
document	2188	27	36
electrifying	2250	0	1
gradation	2202	24	25
image-capturing	2242	3	6
lens	2240	10	1
macro	2243	3	5
motion	2203	25	23
multivalue	2250	0	1
object	2116	90	45
quantization	2221	25	5
radiographic	2243	4	4
retrieval	2219	13	19
roller	2249	0	2
scaling	2245	4	2
signal	1948	105	198
synchronization	2246	2	3
texture	2241	8	2
transfer	2227	18	6
transparency	2244	0	7
ultrasonic	2249	0	2
ultrasound	2247	1	3

We constructed a TF model based on ARM using the training data set. To verify the performance of the constructed model, we used test data set. In general, patent document is not suitable for most data analysis methods such as ARM. Text mining is a good preprocessing tool for transforming patent document into structured data[4]. We use 'tm' R package for text mining[20]. This package provides a text mining framework based on R. R is a language for statistical computing[21]. First, we got DTM using text mining preprocessing. The dimension of DTM was 2251×8512. That is, the numbers of documents and terms were 2251 and 8512 respectively. A element e_{ij} of DTM represents the occurred frequency of term j in document i. There were many meaningless terms in the 8521 terms. They were 'and', 'the', 'for', and so on. We eliminated these terms. We also removed common terms such as 'image', 'video', and 'technology'. Second, we extracted top 30 terms from DTM. Next table shows the terms and their occurred frequency levels.

To apply ARM, we replaced occurred term frequency of DTM by occurred level. For example, if the occurred term frequency was 2 or 3, we determined the level was 'middle'. We knew most frequency levels were 'low'. This was a problem for constructing ARM rules. So, we selected the terms with the number of 'row' levels was respectively small. In above table, the words in bold type were the selected terms. We removed the documents (rows) which have all same levels in ten selected terms. For example, 2250th patent document had binary=low, conversion=low, …, transfer=low. So, we eliminated the document because it was not useful for constructing ARM rules. Third, we found ARM rules in DTM with ten selected terms. The support and confidence were determined as 0.0001 and 0.01 respectively. In this paper, we use 'arules' and 'arulesViz' package of R for our ARM model[22-23]. Finally we got 134,358 ARM rules with support=0.0001 and confidence=0.01. We determined very small values of support and confidence for extracting ARM rules as many as possible. Next, we show four statistics of set of 134,358 rules.

Table 3. Statistics of 134,358 rules

Statistics	Support	Confidence	Lift
Min	0.0009	0.0100	0.1250
Median	0.0045	1.0000	1.0460
Mean	0.0381	0.7554	1.7710
Max	0.9785	1.0000	222.80000

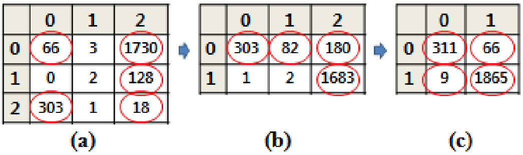
We knew that the average values of support and confidence were large. The average lift value was small relatively. But the maximum value of lift was extremely large. We generated meaningful rules from the set of 134,358 rules. These rules are shown as follow.

We showed top one rule according to each measure for generating ARM rules. First row shows the generated rule with the highest lift value. Second and third rows show the generated rules with the highest confidence and support values. First, in the rule with the highest lift, {(gradation=middle, object=middle)→(quantization=high)}, we found the middle levels of 'gradation' and 'object' were strongly associated with the high level of 'quantization'. Though the occurred frequency of this rule was small, the

Table 4. Generated meaningful rules

{X→Y}	Support	Confidence	Lift
{(gradation=middle,object=middle) → (quantization=high)}	0.0009	1.0000	**222.8000**
{(quantization=high)→ → (signal=low)}	0.0049	**1.0000**	1.3736
{(quantization=low) → (transfer=low)}	**0.9515**	1.0000	0.9994

relationship between them was highly correlated. Also, they were complementary each other strongly. We can use this rule to forecast IVT. Second, the occurred probability of low level of ''signal' given the high level of 'quantization' was 1. That is, the low level of 'signal' was depended too much on the high level of 'quantization' from the rule, {(quantization=high)→(signal=low)}. Third, the low levels of 'quantization' and 'transfer' occurred together. In third rule, {(quantization=low)→(transfer=low)}, we found the low level of 'quantization' and the low level of 'transfer' were needed together for developing IVT. We used SOM as another method for IVT forecasting. For our experiment, 'som' package was used[24]. Next figure shows the clustering result of IVT patent documents. Each cell represents the number of assigned patent documents of a cluster.



	0	1	2
0	66	3	1730
1	0	2	128
2	303	1	18

(a)

	0	1	2
0	303	82	180
1	1	2	1683

(b)

	0	1
0	311	66
1	9	1865

(c)

Fig. 2. SOM results of IVT patent documents clustering – (a) 3×3, (b) 3×2, (c) 2×2

First, we determined the dimension of feature map was 3×3 in (a). Maximum number of clusters was nine. We found the proper number of clusters was five, red circles in (a). Second, we did SOM clustering with 3×2 feature map as (b). In this result, we found the number of clusters was four. Third, we clustered IVT patent documents by 2×2 SOM in (c). We concluded the optimal number of clusters was four. In result (c), we selected a vacant area for IVT forecasting in four clusters. (row, column)=(1,1) and (0,0) were not vacant areas because they had so many patent documents. The cluster, (row, column)=(1,0) had only nine patent documents. We decided the technology of this cluster was an outlier field of IVT. Finally, we also decided (row, column)=(0,1) cluster as a vacant technology area for IVT. This cluster had sixty-six patent documents. This was 2.93% of all training data. To define detailed technology represented by (0,1) cluster, we extracted top ten keywords from the sixty-six patent documents. They were 'distortion', 'luminance', 'signal', 'storage', 'intensity', 'reproduction', 'compress', 'digital', 'quantization', and 'transfer'. We can define the vacant technology of IVT using these ten terms.

Next, to verify the performance of our forecasting method, we used test patent documents. The test data had 1529 patent documents from 2006 to 2010. We found 255 patent documents in the test patent documents. These were 16.68% of all test data. They included ten keywords determined by vacant technology for IVT in the training data. So, we knew that the vacant technology for IVT was increased from training data (1984-2005) to test data (2006-1010). It was because a vacant technology has a chance of increasing in future.

5 Conclusions and Future Works

In this paper, we proposed a forecasting method for IVT vacant technology. We used ARM and SOM for our forecasting model. Patent documents related to IVT were retrieved from USPTO. The patent data had all patents of IVT in U.S., Europe, and China. We forecasted the vacant areas of IVT by constructing forecast models using the patent data. In the ARM, we extracted top keywords from IVT patent documents. These terms were used to find the ARM rules for vacant IVT forecasting. According to the levels of the terms, 'gradation', 'quantization', 'object', 'signal', and 'transfer', we found three ARM rules for IVT forecasting. We used SOM for another vacant TF model of IVT. From large dimension of feature map to small feature map, we searched optimal clustering result of IVT patent documents. We concluded four clusters as optimal clustering. A vacant technology area was decided from the four clusters. We determined a cluster with sixty-six patent documents. This cluster was relatively small but not sparse. We extracted top ten terms from the patent documents belong to the cluster defined as a vacant technology. The percentage of the patents defined vacant area was 2.93% in training patent documents, 1984 to 2005. We knew the percentage of our defined vacant technology in test patents, 2006 to 2010 was 16.68%. So, we verified the forecasting performance of our model.

In this paper, we used patent documents for constructing IVT forecasting models. The other important source containing researched and developed results of IVT was in papers published in journal or conference. One of our future works is to make more accurate TF model of IVT using patents and papers together. A limitation of our work was a shortage of IVT domain experts' support. We needed their knowledge to deploy our experimental results to IVT forecasting.

References

1. Amin, T., Zeytinoglu, M., Guan, L.: Application of Laplacian Mixture Model to Image and Video Retrieval. IEEE Transaction on Multimedia 9(7), 1416–1429 (2007)
2. Okamoto, H., Yasugi, Y., Babaguchi, N., Kitahasui, T.: Video Clustering using Spatio-Temporal Image with Fixed Length. In: IEEE International Conference on Multimedia and Expo., pp. 53–56 (2002)
3. Han, J., Kamber, M.: Data Mining Concepts and Techniques. Morgan Kaufmann (2001)
4. Tseng, Y.H., Lin, C.J., Lin, Y.I.: Text mining techniques for patent analysis. Information Processing & Management 43, 1216–1247 (2007)
5. Madu, C.N., Kuei, C.H., Madu, A.N.: Setting priorities for IT industry in Taiwan-A Delphi study. Long Range Planning 24(5), 105–118 (1991)

6. Mitchell, V.W.: Using Delphi to Forecast in New Technology Industries. Marketing Intelligence & Planning 10(2), 4–9 (1992)
7. Woundenberg, F.: An evaluation of Delphi. Technological Forecasting and Social Change 40, 131–150 (1991)
8. Yun, Y.C., Jeong, G.H., Kim, S.H.: A Delphi technology forecasting approach using a semi-Markov concept. Technological Forecasting and Social Change 40, 273–287 (1991)
9. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207–216 (1993)
10. Hahsler, M., Grun, B., Hornik, K.: arules – A Computational Environment for Mining Association Rules and Frequent Item Sets. Journal of Statistical Software 14(15), 1–25 (2005)
11. Kohonen, T.: Self-Organizing Maps. Springer (2000)
12. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning – Data Mining, Inference, and Prediction. Springer (2001)
13. Metaxiotis, K., Psarras, J.: Expert systems in business: applications and future directions for the operations researcher. Industrial Management & Data Systems 103(5), 361–368 (2003)
14. Yoon, B., Park, Y.: Development of New Technology Forecasting Algorithm: Hybrid Approach for Morphology Analysis and Conjoint Analysis of Patent Information. IEEE Transactions on Engineering Management 54(3), 588–599 (2007)
15. Zhu, D., Porter, A.L.: Automated extraction and visualization of information for technological intelligence and forecasting. Technological Forecastingand Social Change 69, 495–506 (2002)
16. Coates, V., Farooque, M., Klavans, R., Lapid, K., Linstone, H.A., Pistorius, C., Porter, A.L.: On the future of technological forecasting. Technological Forecasting and Social Change 67, 1–17 (2001)
17. Mann, D.L.: Better technology forecasting using systemic innovation methods. Technological Forecasting and Social Change 70, 779–795 (2003)
18. Jun, S., Park, S., Jang, D.: Forecasting Vacant Technology of Patent Analysis System using Self Organizing Map and Matrix Analysis. Journal of the Korea Contents Association 10(2), 462–480 (2010)
19. Jun, S., Uhm, D.: Patent and Statistics, What's the connection? Communications of the Korea Statistical Society 17(2), 205–222 (2010)
20. Feinerer, I., Hornik, K., Meyer, D.: Text Mining Infrastructure in R. Journal of Statistical Software 25(5), 1–54 (2008)
21. R Development Core Team.: R, A language and environment for statistical computing. R Foundation for Statistical Computing (2011), `http://www.R-project.org`
22. Hahsler, M., Buchta, C., Gruen, B., Hornik, K.: Package 'arules'. R-project CRAN (2011)
23. Hahsler, M., Chelluboina, S.: Package 'arulesViz'. R-project CRAN (2011)
24. Yun, J.: Package 'som'. R-project CRAN (2010)