ICIC Express Letters
Part B: Applications
Volume **7**, Number **10**, October **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **2267–2273**

IDENTIFYING POTENTIAL TECHNOLOGY THEMES BASED ON INTERNAL CAPABILITIES USING TOPIC MODELING AND ASSOCIATION RULE MINING

Jusung Kim[1], Woon-Seek Lee[1], Sungchul Choi[2] and Wonchul Seo[1,*]

[1]Division of Systems Management and Engineering
Pukyong National University
45 Yongso-ro, Nam-gu, Busan 608-737, Korea
byulsky12@pukyong.ac.kr; iewslee@pknu.ac.kr; *Corresponding author: wcseo@pknu.ac.kr

[2]Department of Industrial and Management Engineering
Gachon University
1342 Seongnam-dearo, Sujeong-gu, Seongnam-si, Gyeonggi-do 13120, Korea
sc82.choi@gachon.ac.kr

Received March 2016; accepted June 2016

Abstract. *Recently, due to global technology competition, the new technology opportunities discovering in rapidly changing market environment is an essential element in the enterprise. Because these efforts become the basis for the sustainable growth of firms, a number of studies have attempted to suggest systematic methods to discover new technological opportunities. However, these methods have not considered the possibility of realization or only discovered technical opportunities such as the entry of a large set of patent classification codes. Therefore, this paper proposes a method for analyzing internal capabilities in detail and recommending feasible technology opportunities in terms of technology themes which are more specific than general technology classes by utilizing ARM and topic modeling technique. The results of this paper will contribute to saving time, effort and money in the decision-making process of the company's business expansion or changing business area by recommending potential and feasible technology themes.*
Keywords: Technology themes identification, Association rule mining, Topic modeling, Latent Dirichlet allocation

1. **Introduction.** Recently, due to the global technology competition, the new technology opportunities discovering in a rapidly changing market environment is an essential element in the enterprise [1]. Because these efforts become the basis for the sustainable growth of the company [2,3], a number of studies have attempted to suggest systematic methods to discover new technological opportunities: identifying technology opportunities using text mining technique [4], method of detecting new technological opportunities by using subject-action-object (SAO)-based semantic patent analysis and outlier detection [5], proposing a supporting system that uses text mining and morphology analysis in order to understand the trend in the morphology of technology and excavate potential technology opportunities from documents [6], identifying opportunities using keyword-based morphology analysis [7]. However, they have failed to sufficiently consider the internal capabilities of the company. Moreover, they have presented new technology opportunities at the level of patent technology classes [8,9]. Therefore, in this paper, we perform text mining to consider internal capabilities based on the company's patent portfolios, and present a method for recommending new technology themes through quantitative assessment analysis. We expect that the presented method can be a basis for selection of topics for R&D planning of countries, companies, and research institutes. The rest of this paper is organized as follows. Section 2 reviews the related works and Section 3 presents a

method for identifying potential technology themes. A case study is conducted in Section 4 and Section 5 discusses further works and concludes the paper.

2. Groundwork.

2.1. Latent Dirichlet allocation (LDA).

LDA, one of the topic modeling techniques proposed by Blei et al., is a probabilistic model to extract the major topic groups in a word corpus [10]. LDA technique firstly assumes that each of the documents is composed of a set of various topics and secondly, by utilizing Dirichlet distribution, calculates probability that terms appearing in each document are included in a specific topic and finally, by using the drawn result, extracts a set of topics term. LDA diagrammatic model is composed of the number of documents (M), based on the number (N) of words per document word (w), topic (z), the ratio of a topic (θ), parameter (α) for the Dirichlet prior probability of topics distribution per document and parameter (β) for the Dirichlet posterior probability of words distribution per topic [11]. This paper uses the LDA model to obtain information about the technical themes contained in patent documents of the company, and LDA was carried out by the python library Gensim.

2.2. Association rule mining (ARM).

ARM is one of data mining techniques to search for interesting relationships among items in a large database [12]. An association rule stands for the co-occurrence of two items, and indicates that if two items occur together frequently, they have strong association relationships. In this paper, we use two measures, support and confidence, to determine relevance of mined rules. By calculating the probability that each of the items A, B occurs in transaction, we can get the support value that is expressed as $P(A)$, $P(B)$ and means usefulness of mined rule $A \rightarrow B$. And, by calculating the conditional probability that consequent items of the mined rule occur in transactions given that conditional items have already occurred in the same transactions, we can get the confidence value that is expressed as $P(B|A)$ and means the certainty of mined rule $A \rightarrow B$ [4].

3. Method.

This paper proposes a method to recommend potential technology themes based on the internal capabilities (Figure 1).

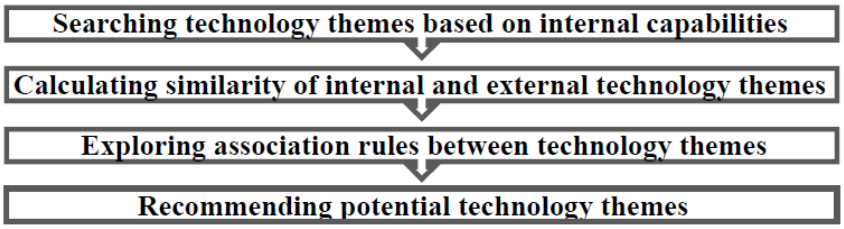

FIGURE 1. Overall procedure of the proposed approach

3.1. Searching technology themes based on internal capabilities.

The first step is to investigate the internal capabilities of the company and to find the technology themes associated with the internal capabilities. First, we collect the patent data of the company to confirm internal capabilities. Second, we perform the LDA after extracting only the summary of the collected patent data and then configuring the corpus, create the patent retrieval query based on internal capabilities to collect patent registered in the United States Patent and Trademark Office (USPTO) by utilizing the terms of constituting the topics. Finally, we perform the new LDA after extracting summary of the patent data collected in the previous step and configuring the new corpus. As a result of the second LDA, we find the topics (Figure 2), and these topics become external technology themes associated with internal capabilities.

```
In [38]: lda.print_topics(10)

Out[38]: [(0,← Topic number
  '0.292*graph + 0.292*minors + 0.153*trees + 0.153*survey + 0.014*system + 0.014*interface + 0.014*user + 0.014*computer + 0.014*response +
0.014*eps'),← weight * term
 (1,
  '0.344*trees + 0.344*graph + 0.031*system + 0.031*interface + 0.031*minors + 0.031*human + 0.031*response + 0.031*user + 0.031*time + 0.03
1*eps'),
```

FIGURE 2. Result of LDA example

3.2. **Calculating similarity of internal and external technology themes.** The second step is to calculate the similarity between internal and external technology themes. In this paper, we utilize similarity calculation function that is another function of the python library Gensim to calculate the similarity. This is a function that calculates the cosine similarity of each document, and does not only simply compare that terms and terms are equal, but also calculate similarity with considering frequency and relevance of terms constituting a document.

3.3. **Exploring association rules between technology themes.** The third step is to explore association rules between technology themes appeared in the first step. It generates the rules based on the co-occurrence between items within a single transaction. In this paper, individual patent and technology themes (topic) correspond to each transaction and item, we can explore the degree of association through the confidence value that is the ratio of co-occurrence frequency with different themes to occurrence frequency of specific technology themes, and we can confirm the validity of the rule through support value that is the ratio of occurrence frequency of specific technology theme to the number of the patent. This process is not only simply exploring between technology themes association rules, also evaluating the strength and validity of the rule.

3.4. **Recommending potential technology themes.** The fourth step is recommending potential technology themes through aggregating data of all the previous steps. In Figure 3, by only considering the above similarity threshold in connection 'internal technology themes ↔ external technology themes', and the above confidence and support threshold in 'Recommend candidate external technology themes', we configure recommending framework. And finally, if the similarity of internal technology themes and potential technology themes is high, we have to carry out the process to delete the connection.

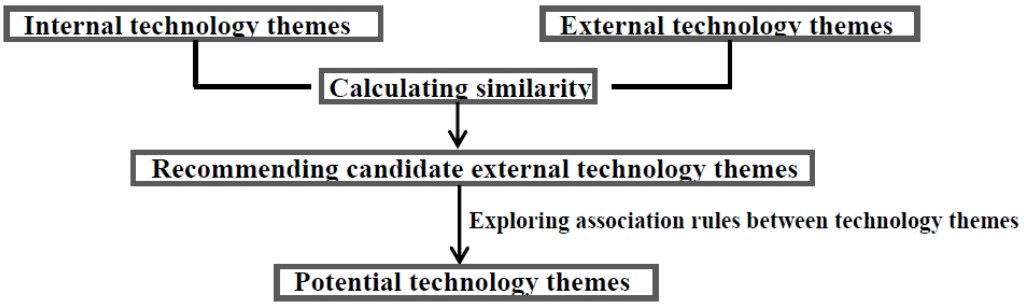

FIGURE 3. A framework for recommending potential technology themes

4. Case Study.

4.1. Collecting data and searching internal and external technology themes.

We perform a case study to explore the practical applicability of the proposed method. We select a target company called 'A.chemistry' and collect its 1,726 patent summary data registered in the USPTO during 2010 and 2015. And then, by performing LDA, we extract 80 topics that are internal technology themes and create the patent retrieval query with combination of top 5 terms which have the highest weight to search external technology themes associated with internal capabilities. Table 1 is an example for creating a patent retrieval query.

We extract a summary of the 9,304 patents data collected in the previous step and configure the new corpus, and then finally perform the new LDA to find new 80 topics associated with the internal capabilities. These topics are external technology themes.

4.2. Discovering similar technology themes.

By utilizing similarity calculation function of Gensim, we calculate similarity of internal and external technology themes (Table 2). We filter over 0.8 cosine similarity entries. In this procedure, if similarity value is high, external technology themes are regarded as internal technology themes. As a result, we obtain the information 32 themes are similar.

Table 1. Part of the patent retrieval query

Topic 10	*power*	*self-diagnosing*	*polylactate*	*charging*	*time*
	fault	*mutant*	*photochromic*	*films*	*propionyl-coa*
Topic 11	*polarizer*	*plate*	*mixed*	*having*	*backlight*
	sheet	*elastic*	*cutting*	*characteristics*	*life*
Patent retrieval query	*((power and self-diagnosing and pla and charging and time) or (polarizer and plate and mixed and having and backlight))*				

Table 2. Discovered similar technology themes

Internal technology themes	*External technology themes*	*Similarity*	*Internal technology themes*	*External technology themes*	*Similarity*
IN00	*EX00*	*0.98*	*IN06*	*EX06*	*0.90*
IN02	*EX02*	*0.97*	*IN41*	*EX30*	*0.89*
IN01	*EX01*	*0.97*	*IN15*	*EX14*	*0.89*
IN78	*EX51*	*0.97*	*IN36*	*EX66*	*0.89*
IN13	*EX12*	*0.96*	*IN04*	*EX04*	*0.88*
IN03	*EX03*	*0.96*	*IN12*	*EX11*	*0.88*
IN14	*EX13*	*0.96*	*IN55*	*EX39*	*0.87*
IN10	*EX09*	*0.95*	*IN58*	*EX41*	*0.87*
IN11	*EX10*	*0.95*	*IN72*	*EX57*	*0.86*
IN26	*EX59*	*0.94*	*IN26*	*EX22*	*0.86*
IN64	*EX44*	*0.93*	*IN29*	*EX24*	*0.83*
IN18	*EX16*	*0.92*	*IN57*	*EX40*	*0.83*
IN23	*EX20*	*0.92*	*IN20*	*EX18*	*0.83*
IN43	*EX68*	*0.92*	*IN19*	*EX17*	*0.82*
IN17	*EX53*	*0.91*	*IN48*	*EX75*	*0.82*
IN37	*EX62*	*0.90*	*IN70*	*EX47*	*0.8*

4.3. **Generating association rules.** We calculate confidence value and support value about the number of 2,528 cases to explore association rules between technology themes. (80 < External technology themes > ∗32 < Technology themes filtered in similarity step > −32 < Self association rules >= 2,528) and then, we filter over 30% confidence value, and 0.5% support value. As a result, we obtain 11 themes that are related to the association rules (Table 3).

4.4. **Identifying potential technology themes and interpretation of result.** In previous step, EX18, 30, 41, 56, 64 are recommended by ARM. However, we know that EX18, 41 are similar to IN20, 58 respectively. So, we have to delete these connections. As a result of all steps, only 3 potential technology themes are recommended based on 7 association rules. Figure 4 is the final result network drawn using Pajek.

Table 3. A result of applying ARM

Conditional themes	*Frequency*	*Consequent themes*	*Frequency*	*Confidence*
EX57	*55*	*EX18*	*1518*	*57.4%*
EX22	*47*	*EX64*	*2148*	*44.7%*
EX53	*57*	*EX64*	*2148*	*41.4%*
EX04	*57*	*EX64*	*2148*	*40.4%*
EX10	*355*	*EX30*	*977*	*36.6%*
EX39	*197*	*Ex41*	*1518*	*36.5%*
EX16	*65*	*EX56*	*551*	*33.8%*
EX09	*55*	*EX41*	*1518*	*32.7%*
EX68	*355*	*EX64*	*977*	*32.5%*
EX22	*47*	*EX41*	*1518*	*31.9%*
EX51	*213*	*EX64*	*2148*	*31.0%*

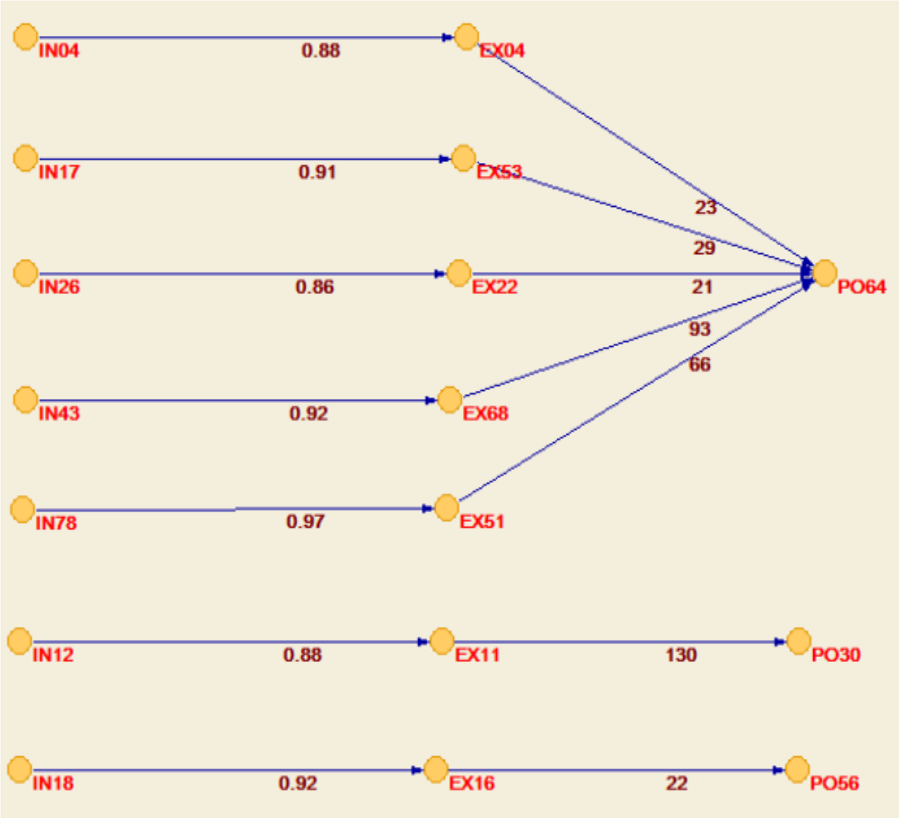

Figure 4. Recommending new technology themes

Table 4. Naming results of technology themes

Technology themes	*Description*
IN04	*Materials for preventing light reflection*
IN17	*System for controlling the water of the fuel cell*
IN26	*Organic polymeric compound*
IN43	*System for estimating the state of the battery cell*
IN78	*Electrochemistry*
IN18	*Thermoelectric Materials*
IN12	*Polarizer*
EX64	*Organic emitters*
EX56	*Luminous display*
EX30	*Display device*

In Figure 4, the values between IN** and EX** are similarity value. EX** is external technology themes that are similar to internal technology themes. The values (23, 29, 21, 93, 66, 130, 22) between EX** and PO** are co-occurrence frequency values. It is used to apply ARM. Consequent themes are written with PO (Potential Themes).

Finally, we need a 'naming process' for network interpretation. For example, IN04 is composed of 10 terms (coated, integral, alchol, with, antireflection, flame, dispersant, module, protective, preparing), so IN04 is named 'Materials for preventing light reflection'. Naming results of every technology theme are the same as Table 4.

Commonality of technology themes that received recommendations as the final result is a technology associated with the display technologies, and generally display technology requires electric technology, polymer technology and so on. So, because internal technology themes have already been associated with electrical, chemical and light, as shown in Table 4, the display technology themes can be recommended to 'A.chemistry' as potential technology themes.

5. **Conclusion.** In this paper, we propose a method to identify and recommend potential technology themes based on internal capabilities of a company. This method can analyze internal capabilities of a company in detail and recommend feasible technology opportunities with technology themes which are more detailed than patent classification code by utilizing ARM and topic modeling technique is significant. However, because this method recommends potential technology themes through association rules between technology themes, there can be already many companies competing in the marketplace and barriers to entry can be very high. Also, recommending unexpected technical themes is hard. However, still because of recommending potential and feasible technology themes, the results of this paper can contribute to saving time, effort and money in the decision-making process of the company's business expansion or changing business area. Moreover, because this method recommends potential technologies as the themes that are clearly more than technology classes, we expect that the presented method can be a basis for selection of topics for R&D planning of countries, companies, and research institutes.

Despite the contribution, further research issues still remain. First, in the process of extracting the topic, meaningless words have been extracted. For example IN04 is composed of 10 terms (coated, integral, alcohol, with, antireflection, flame, dispersant, module, protective, preparing), but term 'with' is a meaningless word. Before extracting the topic, preferentially we calculate the frequency of word that is split into morphological units. And by filtering the meaningless words among morphemes having high occurrence frequency over the threshold, we can remove meaningless words. Because the topic is configured high frequency words in the patents, we think that this method is effective. Second, we need a process to verify the practicality of potential technology themes that

is the result of this methodology in this paper. In carrying out LDA, by classifying the patents by application year, we can know the trend of the technology themes and finally we will find cold topics and hot topics. And then if we find the position of the result topic of this methodology in this paper, it is possible to verify the practicality of the result of this paper.

Acknowledgment. This work was supported by Basic Science Research Program thr-ough the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT&Future Planning (NRF-2014R1A1A1005317).

REFERENCES

[1] R. A. D'Aveni, *Hypercompetition: Managing the Dynamics of Strategic Maneuvering*, Free Press, New York, 1994.

[2] W. J. Abernathy and K. B. Clark, Innovation: Mapping the winds of creative destruction, *Research Policy*, vol.14, no.1, pp.3-22, 1985.

[3] C. Christensen, *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*, Harvard Business Press, Boston, 1997.

[4] W. Seo, J. Yoon, H. Park, B.-Y. Coh, J.-M. Lee and O.-J. Kwon, Product opportunity identification based on internal capabilities using text mining and association rule mining, *Technological Forecasting and Social Change*, vol.105, pp.94-104, 2016.

[5] J. Yoon and K. Kim, Detecting signals of new technological opportunities using semantic patent analysis and outlier detection, *Scientometrics*, vol.90, no.2, pp.445-461, 2012.

[6] B. Yoon, On the development of a technology intelligence tool for identifying technology opportunity, *Expert Systems with Applications*, vol.35, no.1, pp.124-135, 2008.

[7] S. Lee, B. Yoon and Y. Park, An approach to discovering new technology opportunities: Keyword based patent map approach, *Technovation*, vol.29, no.6, pp.481-497, 2009.

[8] A. Ardichvil, R. Cardozo and S. Ray, Theory of entrepreneurship opportunity identification and development, *Journal of Business Venturing*, vol.18, no.1, pp.105-123, 2003.

[9] J. Yoon and K. Kim, Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks, *Scientometrics*, vol.88, no.1, pp.213-228, 2011.

[10] D. M. Blei, A. Y. Ng and M. I. Jordan, Latent dirichlet allocation, *Journal of Machine Learning Research*, vol.3, pp.993-1022, 2003.

[11] J. Fu, N. Liu, C. Hu and X. Zhang, Hot topic classification of microblogging based on cascaded latent dirichlet allocation, *ICIC Express Letters, Part B: Applications*, vol.7, no.3, pp.621-625, 2016.

[12] C. Kim, H. Lee, H. Seol and C. Lee, Identifying core technologies based on technological cross-impacts: An association rule mining (ARM) and analytic network process (ANP) approach, *Expert System with Application*, vol.38, no.10, pp.12559-12564, 2011.