Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2021.013404
Article

Tech Science Press

Identifying Cross Section Technology Application through Chinese Patent Analysis

Ping-Yu Hsu[1], Ming-Shien Cheng[2,*], Chih-Hao Wen[3] and Yen-Huei Ko[1]

[1]Department of Business Administration, National Central University, Jhongli City, 32001, Taiwan
[2]Department of Industrial Engineering and Management, Ming Chi University of Technology, New Taipei City, 24301, Taiwan
[3]Department of Communication Management, Shih-Hsin University, 11604, Taiwan
[*]Corresponding Author: Ming-Shien Cheng. Email: mscheng@mail.mcut.edu.tw
Received: 05 August 2020; Accepted: 15 October 2020

Abstract: Cross-domain technology application is the application of technology from one field to another to create a wide range of application opportunities. To successfully identify emerging technological application cross sections of patent documents is vital to the competitive advantage of companies, and even nations. An automatic process is needed to save precious resources of human experts and exploit huge numbers of patent documents. Chinese patent documents are the source data of our experiment. In this study, an identification algorithm was developed on the basis of a cross-collection mixture model to identify cross section and emerging technology from patents written in Chinese. To verify the algorithm's effectiveness, documents in three transmission-related technology subclasses and one application technology category were collected from WEB-PAT Taiwan. The former subclasses consist of H04B: Transmission; H04L: Transmission of digital information; and H04N: Image communication; and the latter is G06Q: Patents for administration, management, commerce, operation, supervision, or prediction by using data processing systems or methods. Growth rate detection was the most popular approach to forecast emerging technologies, our research defined the growth rate as the difference between the numbers of technology-containing documents published in different time. The emerging technology identified using the proposed method exhibited an average growth rate of 95.08%. By comparison, two benchmark methods identified emerging technology with average growth rates of 9.57% and 51.49%.

Keywords: Cross section analysis; emerging technology identification; Chinese patent analysis; cross-collection mixture model

1 Introduction

Cross-domain technology application uses technology from one field in another, with impacts on human lives and commercial undertakings. Examples include the global positioning system (GPS), radio frequency identification (RFID), light-emitting diode (LED), and financial technology (fintech). The GPS project was launched by the U.S. Department of Defense in 1973 for military use, and became fully operational

in 1995 [1]. This military project has seen significant civilian application in commerce, science, tracking, and surveillance [2]. RFID was developed in the telecommunication field and is now revolutionizing stock tracking and the issuance of goods in business management [3]. LEDs are two-lead semiconductors that emit light when activated. While previously used in illuminating panels [4], they now are often used in communication devices for signal transmission [5]. The emerging field of fintech improves financial activities through technology [6]. The use of smartphones for mobile banking, investing services, and cryptocurrency are examples of technologies that make financial services more accessible to the public [7]. Few have the resources to develop fundamental technologies such as the internet or GPS. However, tracking the adoption and development of source technologies, which are developed within their class, to other application areas is vital to the sustainability of modern companies.

Patent data banks are valuable resources for technological research and documentation. They record more than 80% of all developed technology [8]. Due to its impact on commerce and even national security, the subject of patent management and analysis has been extensively studied. Recent patent analysis studies have detected emerging technology [8–19]. Previous research has either restricted applications to a single class or neglected the classification hierarchy of documents. However, because monitoring cross-domain technological applications is vital to the development of most organizations, a methodology is necessary to analyze patents in different International Patent Classification (IPC) [20] classes to track technological adoption and development among classes.

Since China's accession to the World Trade Organization (WTO) in 2000, the number of patents written in Chinese has increased exponentially [21]. Therefore, technology that can process text documents written in Chinese is essential. The problem in identifying technological terms in such documents is that Chinese characters have multiple meanings. To convey a specific meaning, characters are combined to form words. Analysis of word construction requires extensive linguistic knowledge, so word segmentation is generally conducted using tools developed by professional teams, such as the Chinese Knowledge Information Processing (CKIP) group [22]. However, because technology-related terms tend to be long, terminology identified using such tools tends to be too specific, reducing the possibility of identifying cross-class technology. For example, both "wireless device" and "wireless communication" are related to wireless technology, but they are identified as different words. As a result, a specific approach or human effort must be employed to identify the common technological term "wireless" in both words. This study develops a method to automatically understand terminology to identify cross-class technology applications.

In summary, to successfully identify emerging technological applications across classes of patent documents is crucial to the competitive advantage of corporations, and even countries. An automatic process to solve the language issue is critical to reduce the need for human experts and take advantage of the large numbers of available patent documents. However, to automate this process faces several obstacles:

1. No automatic methodology has been proposed to identify IPC cross-class technology applications.
2. The methodology developed must be able to automatically identify emerging applications of technologies.
3. A novel method is necessary to accommodate the clumsy results of current Chinese word segmentation technologies.

To resolve these issues, this study proposes a methodology based on common and specific theme analysis of patent documents. To avoid litigation, patent documents tend to use different words to describe similar technologies. A keyword-based approach is therefore inadequate. The cross-collection mixture model (CCMM) has been developed to identify common and specific themes [23–27]. Popular concepts among all documents are recorded as common themes, and subconcepts pertaining to particular collections of documents as specific themes. Because technologies developed within their source classes tend to be described with the original concept, they can be identified as common themes. In application

classes, which adopt technology from other classes, the usage of these technologies tends to be annotated with a description of their original applications. As a result, these technological applications can be identified with specific themes.

To verify the effectiveness of the developed method, three subclasses belonging to class H04 were treated as source classes. These were H04B (transmission), H04L (digital information), and H04N (image communication). Class G06Q, belonging to another section, was selected as the application class. G06Q collects patents for administration, management, commerce, operation, supervision, and prediction using data processing systems or methods. The patents were collected from WEBPAT Taiwan [28].

A technology application map was developed to visualize the identified source and application technologies. In this study, three application technologies were identified as emerging technologies. Growth rate detection was the most popular approach to forecast emerging technologies, our research defined the growth rate as the difference between the numbers of technology-containing documents published in different time. The average growth rate of these technologies was 95.08%, whereas those of the technologies identified using two benchmark methods [29,11] were 9.57% and 51.49%.

The remainder of this article is organized as follows. Section 2 reviews the research on the introduction of the IPC, identification of emerging technological terminologies, identification of technological terms in Chinese patents, and cross-collection mixture (theme) models. The proposed model, research design, and methods are detailed in Section 3. The data acquisition process, experiments, visualization of cross-class technology, and identified emerging technologies are discussed in Section 4. Contributions, research limitations of this study, and recommendations for future research are described in Section 5.

2 Literature Review

2.1 Introduction of IPC

The world's most widely used patent classification system, the IPC was established in 1954 with the condition that it be updated every five years [20]. Each classification symbol has the form A01B 1/00. The first letter represents the "section". Combined with the two-digit number represents the "class". The final letter makes up the "subclass" and the following letter indicates the subclass. The subclass is followed by a one-to-three-digit "group" number, an oblique stroke and a number of at least two digits representing a "main group" or "subgroup". A patent examiner assigns classification symbols in a patent application in accordance with classification rules [20]. IPC was last revised in 2016 and consists of 8 sections, 120 classes, 628 subclasses, and 69,000 main items. Tab. 1 shows the main classifications [20].

Table 1: IPC classifications

Section	Classification	Number of main classes
A	Human living necessities	21
B	Operation, transportation	37
C	Chemistry, metallurgy, combination technology	21
D	Textile, paper manufacturing	9
E	Fixed building	8
F	Mechanical engineering, illuminating, heat supply, weapons, blasting	18
G	Physics	14
H	Electrical science	6

2.2 Identification of Emerging Technology Terminologies

Emerging technology shows high potential whose value has not yet been demonstrated or agreed upon by a community of users [9]. Rotolo et al. [19] identified five attributes that feature in the emergence of novel technologies: (i) Radical novelty; (ii) Relatively fast growth; (iii) Coherence; (iv) Prominent impact; and (v) Uncertainty and ambiguity.

Most studies turn to text mining to identify the terminologies representing or symbolizing emerging technologies. This involves identifying n-gram words (is a contiguous sequence of n items from a given sequence of text or speech) [29], keywords with high frequency–inverse document frequency (tf-idf: is a numerical statistic that is intended to reflect how important a word is to a document in a collection or *corpus*) [26], and words frequently used in titles and abstracts [30]. Corrocher et al. [29] proposed a 3-gram method to analyze patent abstracts. This involves collecting 3-gram words from patent abstracts at two different intervals and calculating momentum based on their frequency differences. Words with high momentum are identified as emerging technology terminologies. Momentum has also been used to screen and validate potential emerging technology terminologies in other research [12,14,29,31,32]. Shibata et al. [11] proposed that words with the highest tf-idf values in emerging clusters constitute emerging technology terminologies. Ma and Porter [30] proposed the clustering of keywords to identify emerging topics.

This study utilizes momentums and frequencies to identify terminologies from a set of words included in specific themes, which we explain below. Be reminded that words in specific themes must be further processed due to the inappropriate word phrasing of Chinese word segmentation systems.

2.3 Identification of Technology Terms in Chinese Patents

Research has shown that even among native Chinese speakers, only approximately 75% agreement can be achieved with regard to correct segmentation, and the percentage of agreement decreases as the number of people involved increases [33]. A dictionary-based method enhanced with lexical rules is most commonly used for Chinese word segmentation [33]. A classic approach to applying lexical rules is the maximum (longest term) matching method [34,35], which is based on the assumption that the most meaningful words usually comprise the maximum concatenation of Chinese characters. As a result, the method tends to produce the maximum number of valid words, and it sometimes fails to tokenize hidden subwords. However, because technology terminology is usually long, the technology identified using such tools tends to be highly specific, reducing the possibility of identifying cross section technology terminology in Chinese. Hence, this study proposes a method to further segment words derived from Chinese word segmentation systems.

2.4 Cross-Collection Mixture (Theme) Model

CCMM [27] has been applied to identify potentially relevant terms, and has been used in contextual text mining to identify topic evolution patterns [24] and summarize the history of a theme evolving on a news website. Mei et al. [25] applied this approach to mine spatiotemporal theme patterns on weblogs. They extracted common themes, generated a theme life cycle for each location, and created a theme snapshot for each time period. Mei et al. [36] claimed that this model simultaneously captured the mixture of topics and sentiments. Mei et al. [37] combined this approach with network analysis to summarize topics in text, map a topic onto a network, and discover topical communities within user networks.

Emerging technology application involves transferring popular technology from a source field to an application field to create new applications; therefore, the technology terminology identified in the source field should be associated with common themes, and that in the application field with specific themes related to a particular temporal interval to reflect the freshness of the emerging technology.

3 Research Methodology

In the proposed method, patents are collected from documents in IPC sections. At least one section should provide the source technology, and one should provide the application technology. The collected patent documents should have been published over at least two consecutive years. If the documents are written in Chinese, then a Chinese word segmentation system such as CKIP [22] is used. All segmented words and documents are analyzed using a CCMM. Representative words in a common theme are examined using n-gram methods to identify common technology, which is matched with terminology in specific themes related to the application technology. Subsequently, the common technology and identified cross section technology terminology are shown on a technology map. Terminology with high momentum–appears in years beyond the specified number of years (threshold) is considered emerging. The proposed methodology is diagrammed in Fig. 1.

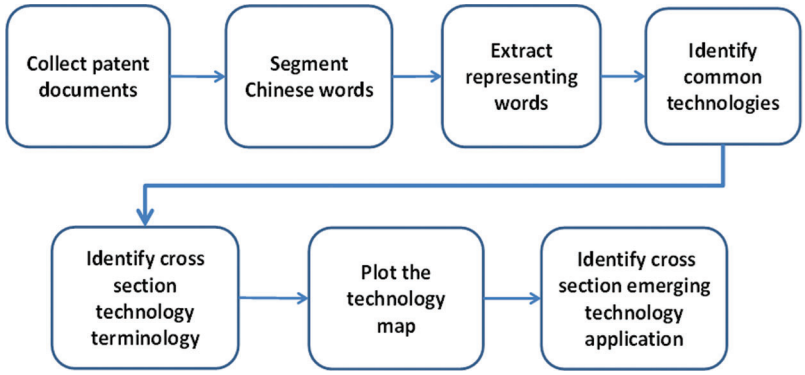

Figure 1: Process of proposed methodology

3.1 Definition of Cross-Collection Mixture (Theme) Model

A theme is a concept derived from a collection of documents and represented by a set of words [23–25,27]. Following the approach of Zhai et al. [17], three theme types were adopted in this research: common, specific, and background. A common theme is a general concept originating from a document collection; a specific theme is a subconcept derived from a subset of documents; and a background theme refers to a group of general words that are closely related to stop words which are filtered out before or after processing of natural language document.

Words highly associated with a common theme are considered popular technologies in the IPC-classified section. Cross section analysis entails the identification of technology that is popular in its own (source) section and emerging in another section. A specific theme within a common theme represents a subconcept derived in a particular year from patent documents published in that year. Terminology for specific themes is considered to denote popular technology in a specific year. Such terminology potentially represents emerging technology in that year.

We define the following based on Zhai's definitions of themes [27].

3.2 Definition 1

a) A theme is a concept shared by a collection of documents. More than one document can share a theme, and a document can address several themes.

b) A set of document collections can address a set of common themes, denoted by $\Theta = \{\theta_1, \ldots, \theta_K\}$.

c) A background theme is a special theme $\theta_B \notin \Theta$ that includes popular stop words in the document collections.

d) Given a collection of documents published at time $t_1, \ldots, t_M$, each common theme $\theta_i \in \Theta$ has a specific theme denoted by θ_{ij}, which represents the subconcept of θ_i derived from documents published at time t_j.

In a theme model, the main purpose of the background theme is to collect and remove words that appear too often as representative words. Words collected under the common theme are those with high probability in the document collections for the entire time frame. A specific theme collects only words with high probability in a certain time period, whose probabilities represent their intensities in a collection.

A document uses a sequence of words from a vocabulary set to describe concepts. Therefore, each document should include several themes. Each theme is associated with a set of words annotated with an intensity probability. Based on Zhai et al.'s research [36], we define the distributions of themes among documents and words among themes.

3.3 Definition 2

a) Given a document d and the set of all addressed common themes $\Theta = \{\theta_1, \ldots, \theta_K\}$, where K is the number of common themes, π_{di} represents the distribution model of d addressing θ_i with the constraint that $\sum_{\theta_i \in \Theta} \pi_{d,i} = 1$.

b) Given a theme θ_i and one of its specific themes θ_{ij}, the vocabulary of all possible words $V = \{w_1, \ldots, w_{|V|}\}$, $p(w_v|\theta_i)$, and $p(w_v|\theta_{ij})$ represent the distribution models of w_v being generated with themes θ_i and θ_{ij}, respectively.

A model of words can be derived based on the distribution models of themes.

3.4 Definition 3

Given a document d published at time t_j, the probability that a user reads a word $w \in d$ is

$$p_{d,j}(w) = \lambda_B p(w|\theta_B) + (1-\lambda_B)\left(\sum_{\theta_i \in \Theta} \pi_{di}\,(1-\lambda_s) p(w|\theta_i) + \lambda_s p(w|\theta_{ij})\right),$$

where λ_B and λ_s are the weights of the inclination of a word to the background and specific themes, respectively.

The values of λ_B and λ_s remain for the user to determine. The higher the value of λ_B, the more likely the words are to be discarded; the higher the value of λ_s, the more sensitive the model is to short-term words. The model is shown in Fig. 2.

Given $C_1, \ldots, C_M$ as a set of documents issued at time $t_1, \ldots, t_M$, and $c(w,d)$ as the count of w in d, an expectation maximum algorithm was designed to maximize the objective function,

$$\log(p(\{C_1, \ldots, C_M\}, \lambda_B, \lambda_S)) = \sum_{\mathrm{j}=1}^{\mathrm{m}} \sum_{\mathrm{d} \in C_j} \sum_{\mathrm{w} \in \mathrm{V}} \mathrm{c(w,d)}\, \log\left(p_{dj}(\mathrm{w})\right). \tag{1}$$

3.5 Parameter Estimation

We discuss the tuning of parameters to maximize data distribution likelihood. Three parameters are fixed before the estimation: $\boldsymbol{\lambda B}$ and $\boldsymbol{\lambda s}$ are manually selected, and $\boldsymbol{p(w|\theta B)}$ is determined as

$$p(w|\theta_B) = \frac{\sum_{j=1}^{m} \sum_{d \in C_j} c(w,d)}{\sum_{j=1}^{m} \sum_{d \in C_j} \sum_{w' \in V} c(w',d)}. \tag{2}$$

An expectation maximization algorithm [38] is used to estimate the remaining parameters by maximizing the data distribution likelihood. The updating rules are as expressed in Eqs. (3)–(8).

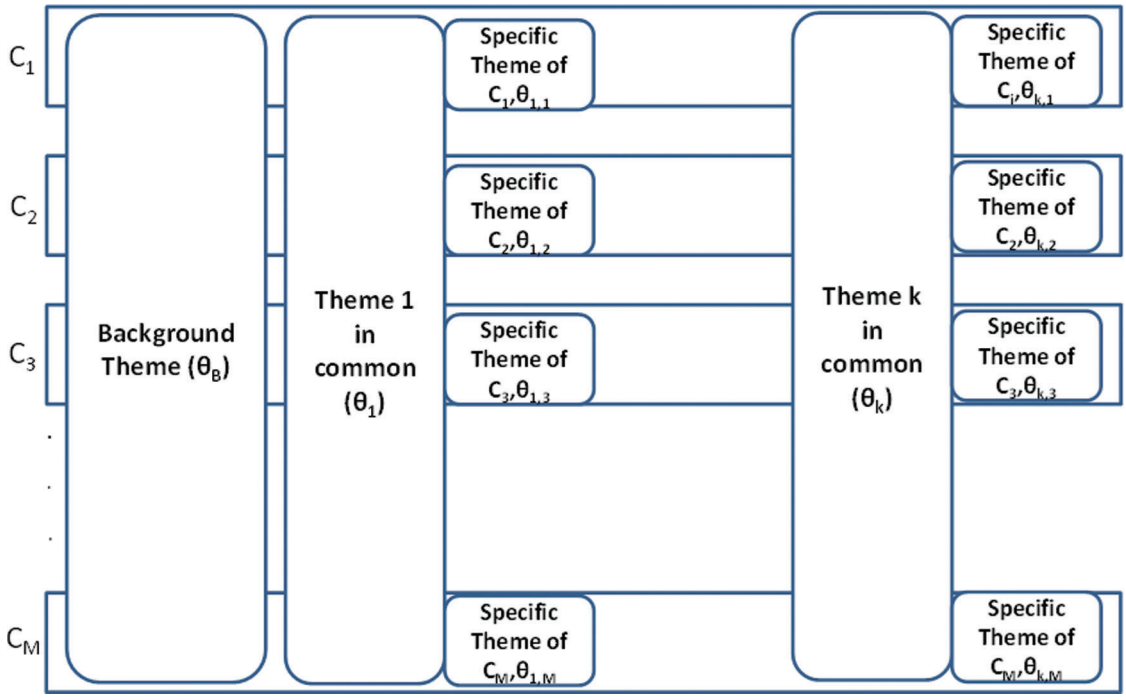

Figure 2: Cross-collection Mixture (Theme) Model

Two hidden variables, Z_{d,w,t_j} and Z_{d,θ_i,w,t_j}, are introduced in the updating rules. Z_{d,w,t_j} represents the theme that w addresses in document d published at time t_j, and Z_{d,θ_i,w,t_j} is a binary value denoting whether w addresses the common theme, but not the specific theme, of θ_i:

$$p(Z_{d,w,t_j}=\theta_i)=\frac{\pi_{d,i}^{(n)}[(1-\lambda_s)p^{(n)}(w|\theta_i)+\lambda_s p^{(n)}(w|\theta_{ij})]}{\sum_{\theta_u\in\theta}\pi_{d,u}^{(n)}[(1-\lambda_s)p^{(n)}(w/\theta_u)+\lambda_s p^{(n)}(w/\theta_{ui})]}, \tag{3}$$

$$p(Z_{d,w,\theta_i,t_j}=com)=\frac{(1-\lambda_s)p^{(n)}(w|\theta_i)}{(1-\lambda_s)p^{(n)}(w|\theta_i)+\lambda_s p^{(n)}(w|\theta_{ij})}, \tag{4}$$

$$p(Z_{d,w,t_j}=\theta_B)=\frac{\lambda_B p(w|\theta_B)}{\lambda_B p^{(n)}(w|\theta_B)+(1-\lambda_B)[\sum_{\theta_u\in\Theta}\pi_{d,u}^{(n)}[(1-\lambda_s)p^{(n)}(w|\theta_u)+\lambda_s p^{(n)}(w|\theta_{uj})]}, \tag{5}$$

$$\pi_{d,i}^{(n+1)}=\frac{\sum_{w\in d}c(w,d)p(Z_{d,w,t_j}=\theta_i)}{\sum_{\theta_u\in\Theta}\sum_{w\in d}c(w,d)p(Z_{d,w,t_j}=\theta_u)}, \tag{6}$$

$$p^{(n+1)}(w|\theta_i)=\frac{\sum_{d\in C}c(w,d)[1-P(Z_{d,w,t_j}=\theta_B)]P(Z_{d,w,t_j}=\theta_i)P(Z_{d,w,\theta_i,t_j}=com)}{\sum_{w'\in V}\sum_{d\in C}c(w',d)[1-P(Z_{d,w',t_j}=\theta_B)]P(Z_{d,w',t_j}=\theta_i)P(Z_{d,w',\theta_i,t_j}=com)}, \tag{7}$$

$$p^{(n+1)}(w|\theta_{ij})=\frac{\sum_{d\in C}c(w,d)[1-P(Z_{d,w,t_j}=\theta_B)]P(Z_{d,w,t_j}=\theta_i)[1-P(Z_{d,w,\theta_i,t_j}=com)]}{\sum_{w'\in V}\sum_{d\in C}c(w',d)[1-P(Z_{d,w',t_j}=\theta_B)]P(Z_{d,w',t_j}=\theta_i)[1-P(Z_{d,w',\theta_i,t_j}=com)]}. \tag{8}$$

We obtained the representative words of common and specific themes defined in Definition 4 to identify popular and potential technology in a given source and application section, respectively.

3.6 Definition 4

a) Given a common theme θ_i, a set of representative words of the common theme is expressed as $W_i = \{\mathrm{w}|\mathrm{w} \in \mathrm{V},\ |\{w_k|w_k \in V, \mathrm{p}(\mathrm{w}|\theta_i) > p(w_k|\theta_i) > 0\}| \geq \varphi_c * |\{w_l|w_l \in V,\ p(w_l|\theta_i) > 0\}|\}$.

b) Given a specific theme θ_{ij}, a set of representative words of the specific theme is expressed as $W_{ij} = \{\mathrm{w}_i|\mathrm{w}_i \in \mathrm{V},\ \left|\{\mathrm{w}_{ik}|\mathrm{w}_{ik} \in V, \mathrm{p}\left(\mathrm{w}|\theta_{ij}\right) > p\left(w_{ik}|\theta_{ij}\right) > 0\}\right| \geq \varphi_s * \left|\{w_{il}|w_{il} \in V,\ p\left(w_{il}|\theta_{ij}\right) > 0\}\right|$,

where φ_c and φ_s are user-provided popularity thresholds.

3.7 Identifying Cross Section Technology Terminology

This study proposes to identify cross section technology terminology using representative words of common and specific themes according to the following observations.

1. Because a specific theme represents a subconcept that is popular during a particular time period, representative words that have high distribution values in a specific theme are candidates for emerging technology terminology.
2. Numerous cross section technology developments integrate popular technology in one section with technology in another section to improve products or services.

An n-gram method is applied to divide Chinese representative words into terms. Terms that appear in sufficient numbers of representative words are considered to denote popular technology. Popular technology terms are then used to identify cross section technology terminology.

Assuming that a word is a sequence of characters, a set of n-gram terms is defined as follows.

3.8 Definition 5

a) A term $t = \{t_1, \ldots, t_u\}$ is a subword of a word $w = \{s_1, \ldots, s_v\}$ and is represented as $t \preceq w$ if $t_1 = s_p$, $t_2 = s_{p+1}, \ldots,$ and $t_u = s_{p+u-1}$.

b) Given a set of words W, a set of n-gram terms of W is given by $W^n = \{w^n|\exists w \epsilon W,\ w^n \preceq w,\ |w^n| = n\}$.

c) Given an n-gram term w^n and a set of n-gram terms W^n, the support (w^n, W^n) is the count of w^n in W^n.

d) Given a common theme θ_i, its representative words W_i, and the set of the top h n-gram terms of a common theme θ_i, $T_i^n = \arg_{\mathrm{top}(h)}\{\mathrm{support}\left(w_i^n, W_i^n\right)\}$.

Cross section technology terminology can be identified using the top (most popular) terms. Representative words of specific application section themes that include popular terms are identified as cross section technology terminology.

3.9 Definition 6

a) Given a specific theme θ_{ij}, its representative words W_{ij}, and the popular n-gram terms of the common theme $l, (1 \leq l \leq K,\ l \neq \mathrm{i}\ )\ T_l^n$, the set of cross section technology terminology structures F is a structure given by $\langle pt,\ it,\ y\rangle$, where pt is the popular term, it is the cross section technology terminology, and y is the year of patent publication: $F = \left\{ < l,\ w_{ij}, j > |w_{ij} \in W_{ij}, t^n \in \cup_{1 \leq l \leq K,\ l \neq i} T_l^n,\ t^n \leq\ w_{ij},\ 1 \leq j \leq M \right\}$.

b) Given a cross section technology terminology structure f, "." is a reference operator. Specifically, $f \cdot pt$, $f \cdot it$, and $f \cdot y$ denote the popular terms, cross section technology terminology, and year, respectively, of technology terminology structure f.

3.10 Identifying Cross Section Emerging Technology Application

Cross section technology terminology that appears in years beyond the specified number of consecutive years is considered emerging. Cross section emerging technology application is defined in Definition 7, and Fig. 3 shows the pseudo-code of an algorithm to identify cross section emerging technology applications.

Algorithm Identifying Cross Section Emerging Technology Application

Input: C \\ Collections of documents,
V: a set of words being segmented by Chinese word segmentation system
I: number of common theme
J: number of specific themes per common theme
a: the designated application class
λ_B, λ_S\\ Parameters of EM
φ_c, φ_s: the threshold of representing words in common and specific themes
n \\ the grams
h \\ top h popular terms
ρ \\ the momentum threshold for cross section emerging technology application

Output:
F // set of structure of cross section technology terminology
E // set of cross section emerging technology application
Utilizing Expectation Maximization to compute $\log(p(C, \lambda_B, \lambda_S))$
For i = 1 to I and $i \neq a$ {

$$W_i = \{w|w \in V,\ |\{w_k|w_k \in V, p(w|\theta_i) > p(w_k|\theta_i) > 0\}| \geq \varphi_c * |\{w_l|w_l \in V, p(w_l|\theta_i) > 0\}|\}$$

$$W_i^n = \{w^n|\exists w \epsilon W_i, w^n \preccurlyeq w, |w^n| = n\}$$

$$T_i^n = \arg_{top(s)}\{support(w_i^n, W_i^n)\}$$

}

For j = 1 to J {

$$W_{aj} = \{w_a|w_a \in V,\ |\{w_{ak}|w_{ak} \in V, p(w|\theta_{aj}) > p(w_{ak}|\theta_{ij}) > 0\}| \geq \varphi_s * |\{w_{al}|w_{al} \in V, p(w_{al}|\theta_{aj}) > 0\}|\}$$

}

$$F = \left\{< l, w_{aj}, j > \middle| w_{aj} \in W_{aj}, t^n \in \cup_{1 \leq l \leq K, l \neq a} T_l^n \ , t^n \preccurlyeq w_{aj}\ , 1 \leq j \leq M\right\}$$

$$E = \{w|\exists f_1, .., f_\rho \in F, f_1.it = .. = f_\rho.it = w, f_k.y = f_1.y + k - 1, 1 \leq k \leq \rho\}$$

Return F, E

Figure 3: Algorithm to identify cross section emerging technology applications

3.11 Definition 7

a) Given a momentum threshold ρ and a set of cross section technology terminology structures F, a set of cross section emerging technologies is given as

$E = \{w|\exists f_1, \ldots, f_\rho \in F, f_1.it = \ldots = f_\rho.it = w, f_k.y = f_1.y + k - 1,\ 1 \leq k \leq \rho\}$.

4 Experiment

4.1 Data Description

Transmission and communication technologies critically affect daily life. Cell phones and other means of wireless communication have given rise to a generation with entirely new information technology consumption behaviors. Therefore, for this study, we chose the subclasses of transmission (H04B), transmission of digital information (H04L), and image communication (H04N) as source classes. To win or maintain competitive advantages, corporations have raced to adapt these technologies and create novel applications. Subclass G06Q includes patents for administration, management, commerce, operation, supervision, and predictions based on data processing systems or methods, and was therefore designated as the application class where the application of cross-class technology should be found.

In total, 1,562 abstracts of patent documents belonging to the four subclasses and published between 2006 and 2011 were collected. Among them, 378, 253, 324, 275, 265, and 67 were published in 2006, 2007, 2008, 2009, 2010, and 2011, respectively. The number of cases published in 2011 is low because only a portion of that year's patents had been submitted when data were being collected for the IPC. Subclasses H04B, H04L, H04N, and G06Q had 339, 448, 396, and 379 cases, respectively, from WEBPAT Taiwan [28].

4.2 Initial Settings of Theme and Word Distribution

Using the four subclasses and six years of data, in CCMM, the number of common themes is the same as the number of patent classes investigated, which is K. The number of specific themes, M, corresponds to the number of years from which the patents were collected.

The theoretical value of π^0_{di} should be between 1 ($d \in \theta_i$) and 0 ($d \notin \theta_i$). Because each document's class was clear in this study, the value should have been either 1 or 0. However, the value 0 prevented the function from updating rules in the CCMM, so we set small nonzero values of $\pi^0_{d,i}$ when $d \notin \theta_i$:

$$\pi^0_{d,i} = \begin{cases} 1-(K-1)*0.01 & d \in \theta_i \\ 0.01 & d \notin \theta_i \end{cases} \quad \sum_{\theta_i \in \Theta} \pi_{d,i} = 1 \quad i = 1, 2, \ldots, K, \tag{9}$$

where

$\sum_{i=1,\ldots,K} \pi^0{}_{d,i} = 1.$

The initial value of the distribution model of each word (w) expressing the common theme θ_i is

$$P^0(w|\theta_i) = (1-\lambda_S)\frac{\sum_{j=1}^{m}\sum_{d\in C_j} \pi_{d,i} c(w,d)}{\sum_{j=1}^{m}\sum_{d\in C_j}\sum_{w'\in V} \pi_{d,i} c(w',d)} \tag{10}$$

The initial value of the distribution model of each word (w) expressing the specific theme θ_{ij} is

$$P^0(w|\theta_{ij}) = \lambda_S \frac{\sum_{d\in C_j} \pi_{d,i} c(w,d)}{\sum_{d\in C_j}\sum_{w'\in V} \pi_{d,i} c(w',d)} \tag{11}$$

4.3 4.3 Setting the Parameters λ_B and λ_S

λ_S and λ_B, respectively, sort words into specific and background themes. Studies have set λ_B as 0.95 [23,25,27]. The value of λ_S affects the distribution of words among specific themes and common themes, in turn affecting the theme distribution among documents. Therefore, we chose the λ_s value that ensured the greatest precision in assigning document themes. In this study, a document d was considered to address theme h if π_{dh} had the highest value among π_{di}, where i = 1,…, K. Studies have reported that

λ_S ranges from 0.2 to 0.8 [23,25,27]. In this study, CCMM performed with the highest precision when λ_S = 0.6, as shown in Fig. 4. This value was chosen for the remainder of the experiment.

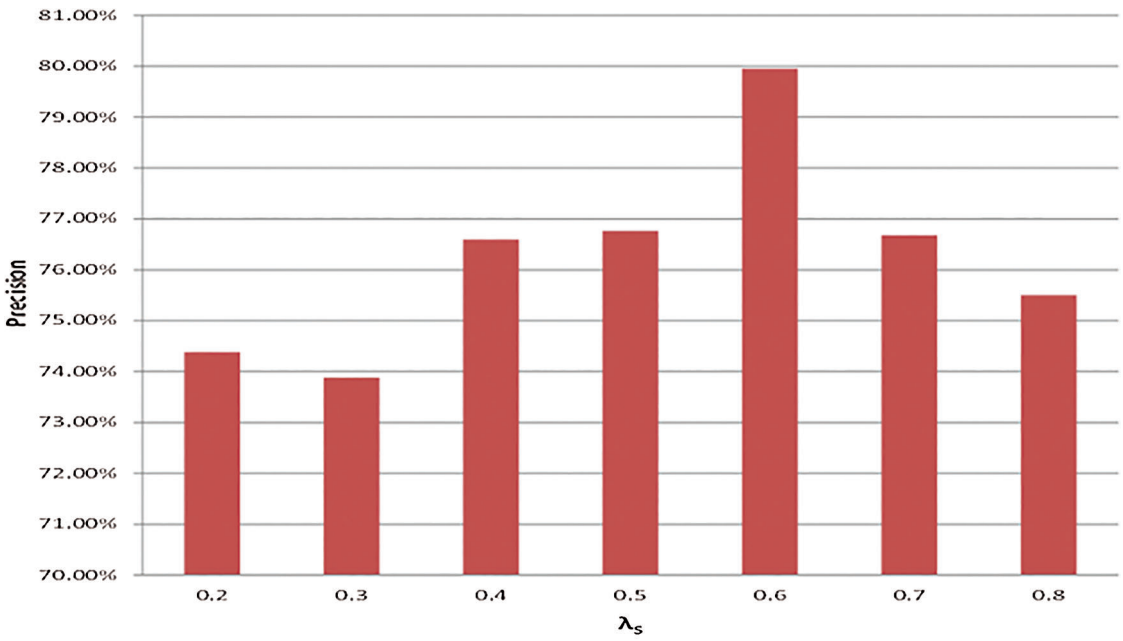

Figure 4: Precision of theme assignment with various λ_S values

Tab. 2 shows the top five representative words of each common theme and their probabilities in the corresponding themes.

Table 2: Top five representative words of each common theme (subclass)

H04B	p(w/θ_1)	H04L	p(w/θ_2)	H04N	p(w/θ_3)	G06Q	p(w/θ_4)
signal(信號)	0.0642848	network(網路)	0.06555026	image(影像)	0.02552709	transaction(交易)	0.0223248
signal(訊號)	0.0350485	packet(封包)	0.03553849	shot(鏡頭)	0.02288817	commodity(商品)	0.0151396
antenna(天線)	0.0198204	message(訊息)	0.02628243	pixel(像素)	0.02180988	member(會員)	0.0142142
frequency(頻率)	0.018826	server(伺服器)	0.02122616	fragment(片段)	0.02123295	order form(訂單)	0.0118526
power(功率)	0.0178861	media(媒體)	0.019644	sensor(感測器)	0.01929434	capacity(產能)	0.0090262

4.4 Evaluating the Quality of Discovered Words

We used the representative words of common themes to identify cross section technology terminology. Therefore, before identifying the terminology, the quality of the identified word distribution among the common themes was evaluated. Paradimitriou et al. [39] proposed an ε-separator index to evaluate the quality of common themes. In a successful model, each theme should be at least 95% distributed among words from the primary set, with the remaining 5% distributed among the other sets. The up-to-5% theme distribution outside the primary set is considered an ϵ-separator index. Tab. 3 shows an ϵ-separator index matrix for all theme combinations. All index values were lower than the 5% threshold. Therefore, the quality of topic distribution was acceptable.

Table 3: Matrix of ϵ-separator Index

ϵ-separator index	H04B	H04L	H04N	G06Q
H04B	0.0000%	0.1584%	0.0204%	0.0002%
H04L	0.0274%	0.0000%	0.0047%	0.0112%
H04N	0.0022%	0.0007%	0.0000%	0.0002%
G06Q	0.0003%	0.0060%	0.0002%	0.0000%

4.5 Identifying Terminology Representing Cross Section Technology

The representative words in themes indicate significant terms in the corresponding section. In this experiment, φ_c and φ_s denote the percentage of words deemed to represent words in common and specific themes, respectively. The weight associated with each word and theme denotes the capability of the word to represent the associated concept. Hence, the higher the weight, the more representative the word. This study discovered as many popular technologies (identified by representing words) as possible. Therefore, the top 50% weighted words were chosen for each common and specific theme.

The n-gram approach was adopted to divide representing words into shorter Chinese words. Most Chinese words are one to four characters long [40], and 84.55% of the representative terms contained fewer than five Chinese characters.

Each term was counted to filter out meaningless or unpopular n-gram terms. Each term was associated with a count tracking the number of representing words including the term. We used a top-h methodology to select the h terms with the highest counts. Fig. 5 shows the total counts versus the value of h. The greatest decrease in average counts occurred when h increased from 3 to 4, so h was set at 3.

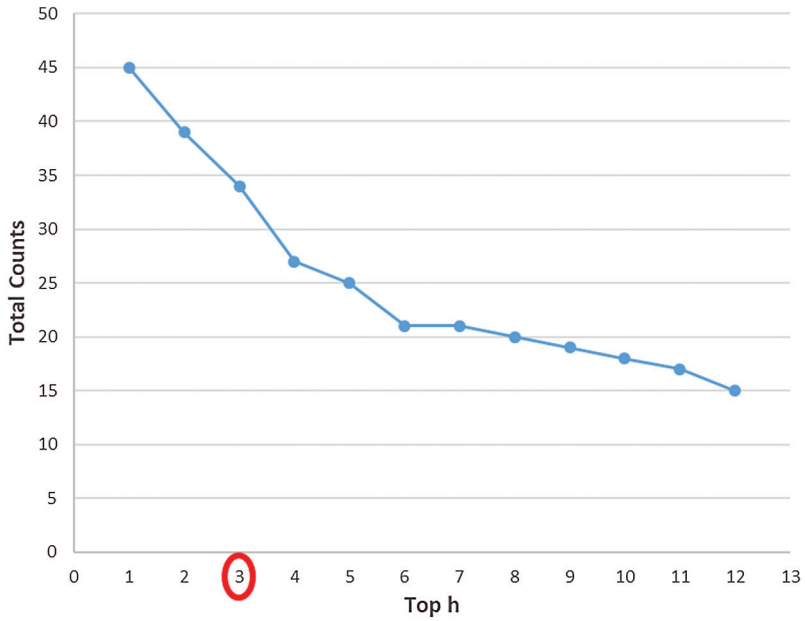

Figure 5: Total counts *vs.* h in three source sections

After the top popular terms were identified (Tab. 4), the cross section technology terminology could be identified. The representative words of specific themes in the application section that included the

popular terms were identified as terminology. Tab. 5 shows the identified cross section technology terminology structures.

Table 4: Top three popular terms for source section

h (Top)	H04B	Support	H04L	Support	H04N	Support
1	Receiving(接收)	11	Wireless(無線)	22	Pixels(像素)	12
2	Wireless(無線)	10	Data(資料)	18	Image(影像)	11
3	Digit(數位)	10	Network(網路)	16	Digit(數位)	8

Table 5: Cross section technology terminology structures

Popular terms	Cross section technology terminology	Year
Wireless(無線)	wireless ID card(無線識別卡)	2008
Wireless(無線)	wireless tag(無線標籤)	2007,2008
Wireless(無線)	RFID(無線射頻)	2007,2008,2009,2010
Digit(數位)	digital right(數位權利)	2006,2007,2008
Digit(數位)	digital database(數位資料庫)	2007
Image(影像)	Image line(影像線)	2008
Network(網路)	network version(網路版型)	2008
Network(網路)	primary network(主要網路)	2011

4.6 Constructing the Cross Section Technology Application Map

The map shows the development of cross section technology through the following features.

1. The source and application section are identified at the top of the map.
2. The representative terms are shown on the left side of the map.
3. In the main part of the figure, the evolution of the cross section technology terminology is shown with the years marked at the top.

Fig. 6 shows the cross section technology map derived from this experiment. The only four popular terms identified were "wireless," "digital," "image," and "network." The cross section technology terminology related to "wireless" comprised "wireless tag" (無線標籤), "wireless ID card" (無線識別卡), and "RFID" (無線射頻). These are all wireless devices that transfer information, and are therefore related to the term "digital," which includes "digital rights" (數位權利) and "digital database" (數位資料庫). The term "image" (影像) includes "image line" (影像線), which is related to "network" (網路). "Network" includes "network version" (網路版型) and "primary network" (主要網路).

4.7 Identifying Cross Section Emerging Technology Terms

Tab. 6 shows eight cross section technology terminologies, of which "wireless tag," "RFID," and "digital rights" exhibited momentum values higher than 2. Depending on the momentum threshold, these three terms may represent emerging technologies [10,11,13]. To gather as many terminologies as possible, all three terms were retained as emerging cross section technology terms.

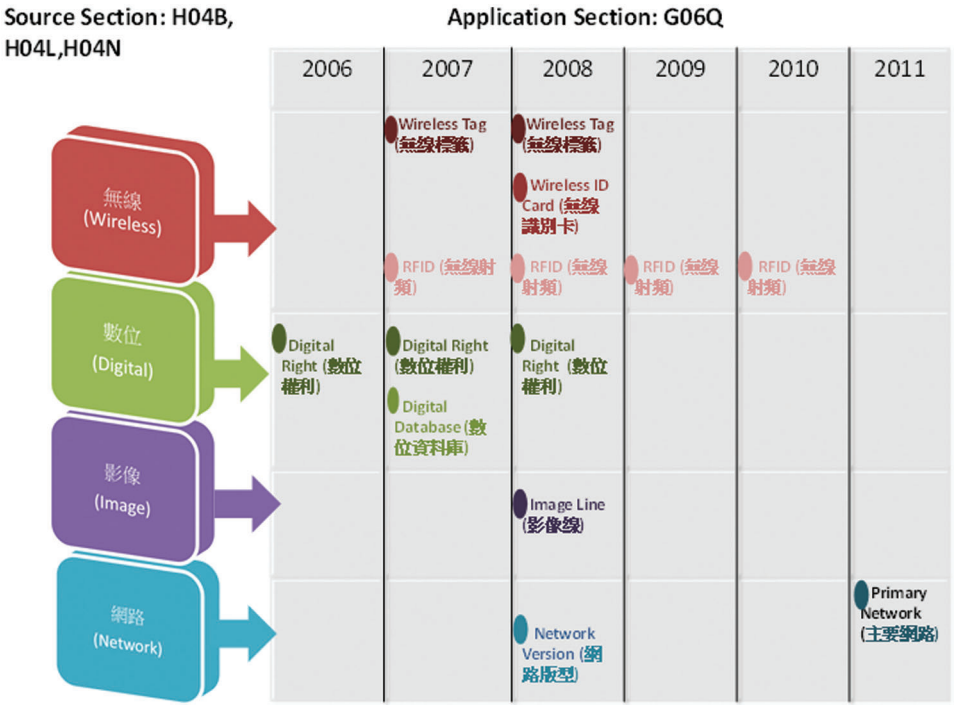

Figure 6: Cross section technology map

Table 6: Performance comparison of three terminology identification methods

	Average growth rate
Our method	95.08%
3-gram method	9.57%
tf-idf method	51.49%

Two other methods have also been proposed to identify emerging terminologies. Corrocher et al. [29] proposed a 3-gram method to identify all 3-gram words with significant frequency in patent abstracts. Shibata et al. [11] used tf-idf to select terminologies with high values. In our control experiment with the tf-idf method, we selected the top five terms from each specific theme. In both of these methods, following the suggestions of Corrocher et al. [29], terms with higher than average growth rates were identified as emerging terminologies. The growth rate is defined as follows: for each term w, the number of documents set (C_1) containing it and published in time t_1 is compared with another documents set (C_2) containing term w and published in time t_2.

$$Growth\_rate(C_1, C_2, w) = \sum_{d \in C_2} c(w,d) - \sum_{d \in C_1} c(w,d) \Big/ \sum_{d \in C_1} c(w,d)$$

In this study, the growth rate of a term w was computed as $Growth\_rate(C_{2005} \cup C_{2006}, C_{2010} \cup C_{2011}, w)$. The growth rate associated with each term was computed as the difference between the numbers of term-containing documents published in 2011 and 2010 and in 2005 and 2006.

5 Conclusion

This study developed a methodology to systemically identify cross-section emerging technological applications between source sections and an application section. Concepts (themes) and representing words in each section were captured by methods revised from CCMM. Besides common themes, CCMM can also capture specific themes developed in each year. Representing words in common themes corresponding to source sections were the technologies that had been developed in that section and that could be adopted in the application section. The representing words of specific themes were the technologies being developed in the application section. These segmented representing words from the source sections were compared against the representing words in the application section to identify cross section technological terminologies. Those with high momentum values were further identified as emerging technologies. The proposed method also generated a technology map to illustrate the adoption of technological terminology in the application section.

To verify the effectiveness of the developed method, four subclasses of patent documents (IPC codes H04B, H04L, H04N, and G06Q) were collected from WEBPAT Taiwan [28]. Three transmission-related technology subclasses, H04B, H04L, and H04N, and one application technology subclass, G06Q, were treated as source and application sections, respectively. The average growth rates of the identified emerging technologies determined from the proposed method, 3-gram approach, and tf–idf methods were 95.08%, 9.57%, and 51.49%, respectively.

This study was explorative and had several limitations. First, the sections covered were limited. Future research should collect patent documents from more varied sections to verify the method's applicability. Second, CCMM is a traditional model, and other refined topic models may be applied to identify word and document distributions. Third, although designed for documents written in Chinese, the proposed model should not be restricted to a specific language.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: We declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. Annex, "*Global Positioning System Standard Positioning Service Performance Standard*," 1995.

[2] Wikipedia, *Global Positioning System*. 2019. [Online]. Available: https://en.wikipedia.org /wiki/Global_Positioning_System.

[3] Wikipedia, *Radio-frequency identification*. 2019. [Online]. Available: https://en.wikipedia.org/wiki/Radio-frequency_identification.

[4] Lighting Research Center, "How is white light made with LEDs?" *Rensselaer Polytechnic Institute*, (cited 12 January, 2019), 2003. [Online]. Available: https://www.lrc.rpi.edu/programs/nlpip/lightinganswers/led/whiteLight.asp.

[5] Wikipedia, *Light-Emitting Diode*. 2016. [Online]. Available: https://en.wikipedia.org/wiki/Light-emitting.

[6] P. Schüffel, "Taming the beast: A scientific definition of fintech," *Journal of Innovation Management*, vol. 4, no. 4, pp. 32–54, 2016.

[7] Wikipedia, "*Financial Technology.* 2017. [Online]. Available: https://en.wikipedia.org/wiki /Financial_technology.

[8] C. Son, Y. Suh, J. Jeon and Y. Park, "Development of a GTM-based patent map for identifying patent vacuums," *Expert Systems with Applications*, vol. 39, no. 3, pp. 2489–2500, 2012.

[9] S. Cozzens, S. Gatchair, J. Kang, K. S. Kim, H. J. Lee *et al.,* "Emerging technologies: Quantitative identification and measurement," *Technology Analysis & Strategic Management*, vol. 22, no. 3, pp. 361–376, 2010.

[10] P. Érdi, K. Makovi, Z. Somogyvári, K. Strandburg, J. Tobochnik *et al.,* "Prediction of emerging technologies based on analysis of the US patent citation network," *Scientometrics*, vol. 95, no. 1, pp. 225–242, 2013.

[11] N. Shibata, Y. Kajikawa, Y. Takeda and K. Matsushima, "Detecting emerging research fronts based on topological measures in citation networks of scientific publications," *Technovation*, vol. 28, no. 11, pp. 758–775, 2008.

[12] M. Bengisu and R. Nekhili, "Forecasting emerging technologies with the aid of science and technology databases," *Technological Forecasting and Social Change*, vol. 73, no. 7, pp. 835–844, 2006.

[13] T. S. Cho and H. Y. Shih, "Patent citation network analysis of core and emerging technologies in Taiwan: 1997-2008," *Scientometrics*, vol. 89, no. 3, pp. 795–811, 2011.

[14] T. U. Daim, G. Rueda, H. Martin and P. Gerdsri, "Forecasting emerging technologies: Use of bibliometrics and patent analysis," *Technological Forecasting and Social Change*, vol. 73, no. 8, pp. 981–1012, 2006.

[15] Y. G. Kim, J. H. Suh and S. C. Park, "Visualization of patent analysis for emerging technology," *Expert Systems with Applications*, vol. 34, no. 3, pp. 1804–1812, 2008.

[16] Y. Ju and S. Y. Sohn, "Patent-based QFD framework development for identification of emerging technologies and related business models: A case of robot technology in Korea," *Technological Forecasting and Social Change*, vol. 94, pp. 44–64, 2015.

[17] C. Lee, O. Kwon, M. Kim and D. Kwon, "Early identification of emerging technologies: A machine learning approach using multiple patent indicators," *Technological Forecasting and Social Change*, vol. 127, pp. 291–303, 2018.

[18] S. Lee, B. Yoon, C. Lee and J. Park, "Business planning based on technological capabilities: Patent analysis for technology-driven roadmapping," *Technological Forecasting and Social Change*, vol. 76, no. 6, pp. 769–786, 2009.

[19] D. Rotolo, D. Hicks and B. R. Martin, "What is an emerging technology?," *Research Policy*, vol. 44, no. 10, pp. 1827–1843, 2015.

[20] Wikipedia, "*International Patent Classification*. 2019. [Online]. Available: https://en.wikipedia.org/wiki/International_Patent_Classification.

[21] H. Kroll, "Exploring the validity of patent applications as an indicator of Chinese competitiveness and market structure," *World Patent Information*, vol. 33, no. 1, pp. 23–33, 2011.

[22] Institute of Information Science Academia Sinica, *Chinese Knowledge Information Processing (CKIP): The Categorical Analysis of Chinese Technical Report*. 2014. [Online]. Available: https://ckip.iis.sinica.edu.tw/.

[23] Q. Mei and C. Zhai, "Discovering evolutionary theme patterns from text: An exploration of temporal text mining," in *Proc. of the Eleventh ACM SIGKDD Int. Conf. on Knowledge Discovery in Data Mining*, Chicago, Illinois, USA, pp. 198–207, 2005.

[24] Q. Mei and C. Zhai, "A mixture model for contextual text mining," in *Proceedings of the 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Philadelphia, PA, USA, pp. 649–655, 2006.

[25] Q. Mei, C. Liu, H. Su and C. Zhai, "A probabilistic approach to spatiotemporal theme pattern mining on weblogs," in *Proc. of the 15th Int. Conf. on World Wide Web*, Edinburgh, Scotland, pp. 533–542, 2006.

[26] S. Morinaga and K. Yamanishi, "Tracking dynamics of topic trends using a finite mixture model," in *Proc. of the Tenth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Seattle, WA, USA, pp. 811–816, 2004.

[27] C. Zhai, A. Velivelli and B. Yu, "A cross-collection mixture model for comparative text mining," in *Proc. of the Tenth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Seattle, WA, USA, pp. 743–748, 2004.

[28] Ministry of Economic Affairs Intellectual Property Office, "Taiwan patent search system." 2009. [Online]. Available: http://twpat-simple.tipo.gov.tw/tipotwoc/tipotwkm.

[29] N. Corrocher, F. Malerba and F. Montobbio, *The Emergence of New Technologies in the ICT Field: Main Actors, Geographical Distribution and Knowledge Sources*. Como, Italian, Department of Economics, University of Insubria, 2003.

[30] J. Ma and A. L. Porter, "Analyzing patent topical information to identify technology pathways and potential opportunities," *Scientometrics*, vol. 102, no. 1, pp. 811–827, 2015.

[31] A. Ávila-Robinson and K. Miyazaki, "Dynamics of scientific knowledge bases as proxies for discerning technological emergence—The case of MEMS/NEMS technologies," *Technological Forecasting and Social Change*, vol. 80, no. 6, pp. 1071–1084, 2013.

[32] D. K. Robinson, L. Huang, Y. Guo and A. L. Porter, "Forecasting innovation pathways (FIP) for new and emerging science and technologies," *Technological Forecasting and Social Change*, vol. 80, no. 2, pp. 267–285, 2013.

[33] Y. C. Liu and C. W. Lin, "A new method to compose long unknown Chinese keywords," *Journal of Information Science*, vol. 38, no. 4, pp. 366–382, 2012.

[34] Q. X. Lin, C. H. Chang, C. L. Chen, L. C. Yu, C. H. Wu *et al.,* "A simple and effective closed test for Chinese word segmentation based on sequence labeling," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 15, pp. 3–4, 2010.

[35] P. K. Wong and C. Chan, "Chinese word segmentation based on maximum matching and word binding force," in *Proc. of the 16th Conf. on Computational Linguistics*, Taipei, Taiwan, Association for Computational Linguistics, vol. 1, pp. 200–203, 1996.

[36] Q. Mei, X. Ling, M. Wondra, H. Su and C. Zhai, "Topic sentiment mixture: Modeling facets and opinions in weblogs," in *Proc. of the 16th Int. Conf. on World Wide Web*, Banff, Alberta, Canada, pp. 171–180, 2007.

[37] Q. Mei, D. Cai, D. Zhang and C. Zhai, "Topic modeling with network regularization," in *Proc. of the 17th Int. Conf. on World Wide Web*, Beijing, China, pp. 101–110, 2008.

[38] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[39] C. H. Papadimitriou, H. Tamaki, P. Raghavan and S. Vempala, "Latent semantic indexing: A probabilistic analysis," in *Proc. of the Seventeenth ACM SIGACT-SIGMOD-SIGART Sym. on Principles of Database Systems*, Seattle, Washington, USA, pp. 159–168, 1998.

[40] J. S. Chang, Y. C. Lin and K. Y. Su, "Automatic construction of a Chinese electronic dictionary," in *Proc. of the Third Workshop on Very Large Corpora*, Cambridge, Massachusetts, USA, 1995.