Journal of The Korea Society of Computer and Information
Vol. 24 No. 6, pp. 89-97, June 2019
www.ksci.re.kr
https://doi.org/10.9708/jksci.2019.24.06.089

# A Study on the Integration Between Smart Mobility Technology and Information Communication Technology (ICT) Using Patent Analysis

Khaled Sulaiman Khalfan Sulaiman Alkaabi*, Jiwon Yu*

## Abstract

This study proposes a method for investigating current patents related to information communication technology and smart mobility to provide insights into future technology trends. The method is based on text mining clustering analysis. The method consists of two stages, which are data preparation and clustering analysis, respectively. In the first stage, tokenizing, filtering, stemming, and feature selection are implemented to transform the data into a usable format (structured data) and to extract useful information for the next stage. In the second stage, the structured data is partitioned into groups. The K-medoids algorithm is selected over the K-means algorithm for this analysis owing to its advantages in dealing with noise and outliers. The results of the analysis indicate that most current patents focus mainly on smart connectivity and smart guide systems, which play a major role in the development of smart mobility.

▸Keyword: Information Communication Technology (ICT), Smart Mobility, Text Mining, Clustering, Patent Analysis

## I. Introduction

The world is changing rapidly owing to the advanced development of artificial intelligence (AI), big data, and the internet of things (IoT). The shift from conventional automation technology to smart communication technology is ushering in a new era known as the Fourth Industrial Revolution. An example of this advancement is Industry 4.0, a German strategic initiative that focuses on the transition from conventional centralized factories to smart factories, in which the interaction and communication between machines are decentralized based on information communication technology (ICT). Moreover, owing to the rapid development of all types of smart devices, such as smart phones, their applications, and sensors based on AI, big data, and the IoT, the capabilities of these devices are becoming tremendous. They can therefore be used to collect and analyze various amount of data that can aid in the transition towards the Fourth Industrial Revolution.

These advanced developments are driven mostly by economic and environmental factors. In terms of economic factors, due to the economic recession, consumer behavior has shifted towards the so-called sharing economy, which has changed all aspects of consumers' lives. Now, consumers can buy products in just a few seconds by using smart phone applications connected to the internet rather than going to physical shops to buy their desired products. In addition, the total number of people using smart phones has increased rapidly, and transactions using applications on smart devices have increased and are expected to further expand in the near future. This expansion will create new services that can

• First Author: Khaled Sulaiman Khalfan Sulaiman Alkaabi, Corresponding Author: Jiwon Yu
*Khaled Sulaiman Khalfan Sulaiman Alkaabi (khaled12@korea.ac.kr), Dpt. of Industrial Management Engineering, Korea University
*Jiwon Yu (vermouth28@korea.ac.kr), Dpt. of Industrial Management Engineering, Korea University
• Received: 2019. 05. 09, Revised: 2019. 06. 08, Accepted: 2019. 06. 10.

bring new business opportunities. By 2027, it has been projected that online sales will surpass $1 trillion, in the opinion of FIT Consulting Inc., U.S. [1].

In terms of the environment, according to the United Nations, those living in urban areas currently make up 55% of the global population but other forecasts project 68% by 2050 [2]. As a consequence of this growth, traffic congestion, road accidents, and emissions have become critical issues. For instance, the cost of congestion and road accidents associated with urbanization issues in the US reached $305 billion in 2017 [3]. Additionally, the future of mobility is moving toward the three objectives of zero emissions, zero accidents, and zero ownership [4].

The idea of developing smart sustainable cities has become a major concern of governments worldwide owing to the apparent benefits that can be obtained from implementing this concept in practice. According to FG-SSG, a smart sustainable city is defined as "an innovative city that uses ICTs and other means to improve quality of life, efficiency of urban operation and services, and competitiveness, while ensuring that it meets the needs of present and future generations with respect to economic, social and environmental aspects" [5].

Governments are seeking to develop and establish ICT infrastructures to achieve many objectives, such as cost reductions, safety, security, reduced emissions, and so on. Many efforts and initiatives have been undertaken by companies and governments globally to develop suitable technology and infrastructure for the future of smart mobility based on ICT to mitigate the problems associated with the increasing urban population. Thus, it is necessary to analyze the current technology trends related to smart mobility.

To conduct such an analysis, we collect US patent data related to smart mobility. This data set is unstructured, and a modification process is required to convert it to structured data. Thus, we use text mining techniques to convert our collected data into structured data for further analysis. Data pre-processing and feature extraction are the two main stages when considering the text mining process. Data pre-processing helps to convert the data into structured data, and feature extraction helps to extract valuable information from text. Additionally, the data is assigned into groups using a clustering algorithm, and we use a Silhouette measure index to determine the optimal number of clusters.

The rest of this paper is organized as follows. Section 2 provides an overview of the related work, methodology, patent analysis, text mining, and clustering algorithm. The numerical analysis is described in section 3. Finally, section 4 concludes the study.

## II. Preliminaries

### 1. Related work

Regarding the work associated with smart cities and by looking at the related literature, it turns out that there is no single definition to define a smart city, however, there are other definitions that overlap such as "digital city", "intelligent city", "creative city", "smart community", "knowledge city" and so on [6] - [9]. Moreover, the smart city concept is broad and segmented into several entities as follows: smart living, smart environment, smart people, smart economy, smart governance and smart mobility [8]. Additionally, a study is conducted for analytical purposes between five government initiatives across the globe in order to define the failure factors as well as success points of those smart sustainable cities [12]. Additionally, a literature review about smart cities is conducted in terms of computer science and information technology by analyzing scientific publications in order to convey the big picture and analyze the trends about the domain of smart cities by using data analysis techniques [11].

Moreover, a study is conducted in order to assess the smart mobility of Cagliari city, a comparison between Cagliari and other international cities is implemented in terms of public transport, alternative mobility options and technological mobility services with the aim of developing Cagliari`s smart mobility [10].

On the other hand, our work is different than the afore-mentioned researches since, in this research, patent data is collected from Wipson Database in order to investigate current market trends related to information communication technology (ICT) and smart mobility in order to provide insights into future technology trends.

### 2. Methodology and research framework

This study focuses on a clustering analysis of patent data. We use data from the abstract sections of patents as inputs. The analytical procedure includes the following

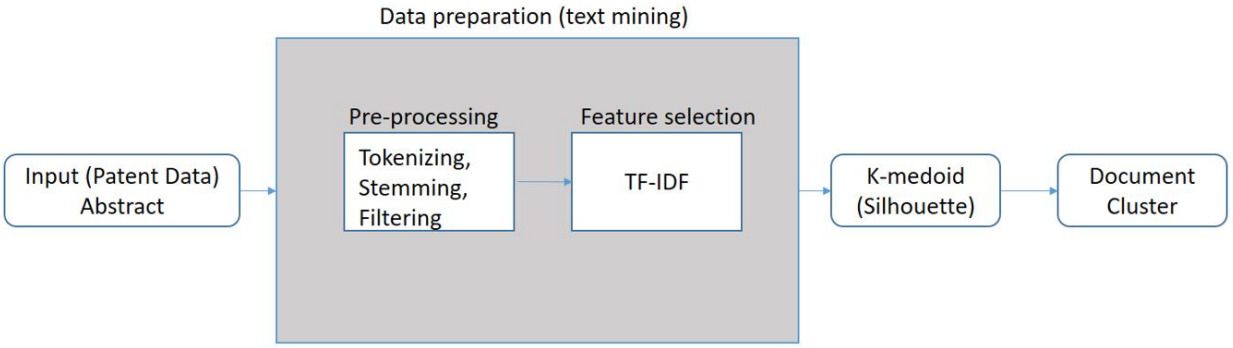


Fig. 1. Methodology

elements (as shown in Figure 1):

- Obtaining the abstracts of patent documents from the patent documents themselves;
- Carrying out the pre-processing phase, which includes tokenizing, filtering, and stemming to obtain a term-document matrix that identifies the frequency of occurrence of concrete words in each chosen document;
- Realizing the feature selection (information extraction) phase using the term frequency-inverse document frequency (TF-IDF) method to obtain a superbly weighted matrix;
- Performing a clustering analysis of the matrix obtained in the previous step using the K-medoids method;
- Analyzing the Silhouette width measure index to define the reliability of the outcomes of the clustering procedure.

## 3. Patent analysis

Patents are the form of intellectual property on innovations and inventories produced by a research group in any type of organization. Patent rights suppose that the owner of a patent obtains the exclusive right to use and profit from the intellectual property and are assumed to "be valid, clearly defined in scope, clearly defined in the statutory term, and single patent should cover an entire product or process innovation" [13]. The patent mechanism is meant to encourage the economic and technological development of modern society, creating incentives for business actors to work actively in the sphere of research and development. Without the patent system, investors in the business environment might not be motivated to develop innovations in different aspects of public life because their interests and costs would not be protected [14]. Additionally, without the patent system for protecting intellectual property, investors and research groups would have difficulty identifying the returns to their research costs and protecting their interests in the process of publicly applying the results of their research. Here, the patent data protection mechanism is important because it protects investors' interests over the long term and serves as a main guarantee and source of profits from research [15]. Without the proper protection of investors' rights, the level of motivation for research and innovation would be significantly lower.

To better understand the potential impact of patent data on the development of modern society, we must compare the benefits and costs of this intellectual property protection system. The patent system can contribute to the development of a significant monopoly when a single player in society obtains relevant innovative technologies and uses them to gain a competitive advantage in the market and control other companies [16]. However, experts usually describe the patent system of intellectual property protection as a tradeoff between short-term costs and long-term benefits. In the short term, certain costs are connected with an investors' monopolistic ownership of the innovative technologies protected by the patent [17], whereas, in the long term, the patent system of intellectual property protection allows society to develop innovations and move technology and the economy forward [17].

From this perspective, the patent system is crucial for the survival and development of human society in general. The minimization of short-term costs caused by the granting of monopolistic rights over technologies is an important task for any society, and this task can be addressed by developing a socially responsible model of business development within the society. Nevertheless, in modern society, there is no more effective alternative to patent rights.

## 4. Text mining

Text mining is a process of applying algorithms from machine learning and statistics with the aim of finding useful information in text [18]. Its main objective is to

find useful information without duplication from different sources that offer synonymous meanings. There is a range of standard tasks in the text mining process "information retrieval, document summarization, text clustering, entity relation modeling, text categorization, information extraction, and sentiment analysis" [18]. Altogether, understanding the procedures that should be applied to text mining is critical for successfully extracting important information from the text. The effective implementation of text mining involves two stages, data preparation and data analysis, and each of these processes can follow several strategies.

Data pre-processing is a text mining technique in which raw data is transformed into an easily understood format to ascertain whether the data convey their meaning to the subjects or audience for which they are intended [19]. Real world data is often incomplete and tend to constitute many errors [20]. Thus, pre-processing is required to align the inconsistencies so that inherent trends may be revealed. In the data preparation or pre-processing stage, different methodologies, namely, tokenizing, filtering [18], and information extraction [21], can be applied to the data.

Tokenizing is an active data pre-processing step that helps to dismantle the text into separate elements, which are regarded as tokens, and changes capital letters into lowercase [22] - [24]. Filtering is an essential part of text mining that helps to remove trivial tokens that are frequently found in the corpus but do not impact the text context [23], [25].

Additionally, stemming is a form of data pre-processing in text mining in which text is reduced to its root as a result of removing an inflection. The purpose of this step is to avoid misinterpretation. In this process, the features that are considered unnecessary are dropped in preparation for assigning meaning to the text so that the correct interpretation may be attained. Stemming occurs with reference to the context in which data is intended to be used [26].

Moreover, feature selection is a text mining process that involves reducing the inputs used for processing and analysis. Its primary intention is to locate the most meaningful inputs in a large text so that the desired meaning is conveyed in the final analysis stage [27]. Only useful information and features are extracted from the existing data to give the desired meaning. One of the tools that can be used for feature selection is the TF-IDF method.

The TF-IDF method generates a numerical statistic that reflects the importance of a word in an entire document that constitutes an enormous set of structured texts. This measure is the weighing factor during searches focused on information retrieval, and its value is directly proportional to the number of times that a particular word appears in a given document [28]. Thus, necessary adjustments may be made to the course of analysis in reference to the objective intended to be achieved.

The purpose of the TF-IDF method is attaining an effective balance in the distribution of words in a text during analysis to conform to the intention that is meant to be served. The effective use of TF-IDF alludes to the knowledge of the frequency with which a word is used in a large set of structured text so that a measure of the significance of the given term in the entire set may be determined [29]. Thus, the value of the text may be determined with the knowledge of the essential word components.

## 5. Clustering algorithm

The second phase of text mining is more complex and integrates a wide range of solutions and methods that can be applied, including text clustering and classification [21]. Clustering is one of the most important approaches to the resolution of the issue of text mining using statistical methods. In addition, a variety of methods are used to solve the issue of text data clustering and analysis, among which the K-means and K-medoids methods are the most valuable.

The objective of the K-means method is to identify data clusters that can be utilized to structure information and divide sections of text into groups [30]. Currently, the K-means method is considered one of the most effective text mining algorithms with the aim of unsupervised learning [31]. The benefits of the K-means approach include its "small computational complexity, high efficiency for large datasets, and high linearity of time complex" [32].

That said, the K-means method is not the only effective solution for data clustering in text mining. The K-medoids method is an adaptation of the K-means method that is supposed to minimize the squared errors in clusters when processing big data [33]. The main difference between this algorithm and the solution used in the K-means method is that, with the K-medoids method, the researcher focuses on searching the medoids rather than the centers of each data cluster. The following

algorithm is utilized for the K-medoids method of text mining: “K elements are randomly identified to be the initial cluster medoids, each data element is assigned to the cluster associated with the closest medoid, positions of medoids are recalculated to reduce the squared errors sum, and this procedure is repeated until the medoids become fixed” [34]. Clearly, the K-medoids algorithm has fewer steps. Moreover, the medoids are not initially fixed and, thus, can be changed according to changes in the amount of data for analysis, which is this method’s main advantage over the K-means method. Additionally, the performance of K-medoids algorithm is better than that of the K-means algorithm in terms of the initialization time and sensitivity to outliers and noise [35] - [36]. Thus, this study uses the K-medoids algorithm rather than the K-means algorithm.

In the process of text mining with the help of clustering techniques, assessing the quality of the achieved cluster structure is an important part of the analysis. Here, a variety of solutions are available. Among them, one of the most valuable and useful tools is the Silhouette width measure index. This method is applied to analyze the effectiveness of the clustering procedure based on an assessment of the widths inside and between each cluster. The advantage of this method is that it can be used to analyze not only the quality of the single cluster identification but also that of the structure of the group of clusters in the process of text mining [37]. It is a very simple cluster quality assessment procedure that can be utilized in any clustering algorithm [38].

The main elements of the Silhouette width measure index algorithm of cluster quality assessment should be mentioned as well. The assessment of the Silhouette width measure index involves defining the average dissimilarity of points in a single cluster and comparing this result with the minimum average dissimilarity to all data points in another cluster [39]. In this way, the Silhouette width measure index can range from -1 to 1 [40]. A Silhouette width measure index of 1 indicates that the focal element is assigned to the right cluster, whereas an index of -1 indicates that the focal element could be assigned to the opposite cluster without any risk of cluster quality loss. In this way, the Silhouette width measure index is used to assess the effectiveness of the clustering procedure and the right locations of each data element in the clusters.

# III. Numerical analysis

## 1. Scope of analysis

The scope of this study is investigating and analyzing current patents related to smart mobility technology to provide insights into future technology trends. For this purpose, a set of open and registered US patents from 1990 to 2017 were collected from the Wipson Database.

## 2. Analytical results

### 2.1 Text mining

For the analysis, text mining techniques (e.g., pre-processing and feature selection techniques) are applied to convert the text, which is regarded as unstructured data, to structured data. Table 1 shows an example of the text when it is transformed into structured data. The first stage of text mining is pre-processing, which consists of tokenizing, stemming, and filtering. Regarding the tokenizing, the text is broken down into separate elements that are regarded as tokens, and the capital letters are changed into lower case letters, as shown in Table 1. Additionally, the base form of the

Table 1. An example of the results obtained by applying text mining techniques

| Terms | Doc1 | Doc2 | Doc3 | Doc4 | Doc5 | Doc6 | Doc7 | Doc8 | Doc9 | Doc10 |
|---|---|---|---|---|---|---|---|---|---|---|
| addit | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| beacon | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| bicycl | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| bluetooth | 1 | 0 | 1 | 0 | 1 | 2 | 0 | 0 | 1 | 1 |
| brake | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| chang | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| communic | 1 | 2 | 0 | 1 | 1 | 1 | 0 | 0 | 3 | 0 |
| coupl | 1 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| design | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| energi | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |

tokens are used, as word endings or inflections are removed during the stemming phase. Then the lesser tokens, which are frequently found in the corpus, are removed in the filtering stage. In addition, feature selection is applied by using TF-IDF to gather useful information from the text. The above procedure is continuously applied until we extract useful information from the text. Moreover, the number of appearances of each term in each document is shown in Table 1. For instance, the terms such as “bluetooth” and “communic” appeared frequently in the documents whereas the term “addit” appeared less frequently in the text, which is regarded as lesser tokens.

### 2.2 Clustering

Regarding clustering, K-medoids algorithm is implemented to partition the data set into groups. Additionally, to select the optimal number of clusters, Silhouette width measure index is determined. The optimal number of clusters is two, as shown in Figure 2. When there are two clusters, the average width rises to be 0.15, which is regarded as the highest average width. Whereas, the average Silhouette width drops between zero and 0.05 when the number of clusters is more than two. Thus, we conclude that the optimal result is yielded when there are two clusters.

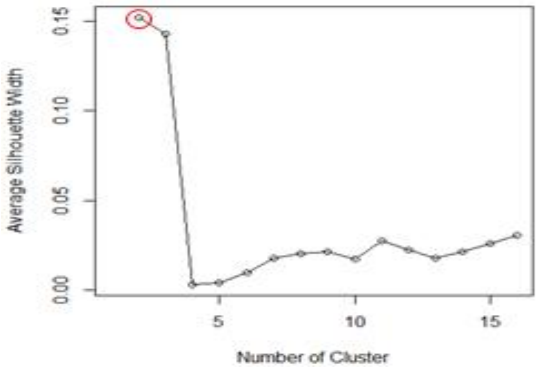


Fig. 2. The optimal number of clusters based on Silhouette width

### 2.3 Keyword extraction from each cluster

The top ten keywords are extracted from each group, as shown in Table 2. Additionally, Table 2 shows that the technologies related to ICT and smart mobility are divided in to two areas: smart connectivity systems (Cluster 1) and smart guide systems (Cluster 2).

Table 2. Keywords for each cluster.

| Cluster 1 | Cluster2 |
|---|---|
| RFID | RFID |
| BLE | BLE |
| Wireless | FOB |
| Bluetooth | Tag |
| Beacon | Magnetic |
| Network | Barcode |
| Signal | Chip |
| Radio | Antenna |
| Location | Controller |
| Block | Bluetooth |

#### 2.3.1 The first cluster: smart connectivity systems.

Most of the patents in the first cluster are based on wireless technology, such as Bluetooth Low Energy (BLE) technology. These patents focus on data collection and information sharing as well as methods for transmitting and receiving data in wireless communication systems to improve wireless device performance and manage connectivity between devices. Moreover, a BLE mesh network technology was developed to enable communication between an enormous number of devices for the purpose of information sharing and enhance interactions between devices, enabling decentralized communication. Additionally, systems and methods based on BLE technology are also proposed to provide accurate location information in three-dimensional space.

#### 2.3.2 The second cluster: smart guide systems.

The systems and the devices found in this cluster are based on radio-frequency identification (RFID) and BLE technology. Regarding RFID technology, most of the proposed patents relate to route and destination guide applications that support and help visually impaired persons reach their destinations easily using RFID readers and tags placed in Braille blocks. These devices provide a voice alarm or vibrator to notify and guide the user to safely reach the target destination. Moreover, the proposed RFID technology can help the user be aware of the surrounding environment to avoid accidents that may occur due to nearby obstacles by sending notifications and providing alternative routes to avoid accidents. Additionally, BLE technologies are proposed to track patients, collect data for the surrounding area, and to monitor the user’s health status. Moreover, an emergency response method on a network of BLE beacons that includes identification information and emergency response alerts are also proposed.

### 2.3.3 Patent trends in the US market.

Regarding the US market, the countries that contribute to smart connectivity systems and smart guide systems are as follows: Republic of Korea, China, Japan, and other countries worldwide (as shown in Figures 3 and 4). Moreover, the number of patents provided by each country are shown in Table 3. Additionally, in both groups, the country that provides the most contributions is the US, followed by the Republic of Korea. Moreover, more contributions relate to smart connectivity systems than to smart guide systems, as the total number of patents for these groups are 348 and 268, respectively. Furthermore, both technologies (smart connectivity systems and smart guide systems) are actively being developed and have a high potential to become among high priorities in the technology development sector.

Table 3. The countries contributing to the US market

| | Cluster1 | Cluster2 |
|---|---|---|
| US | 222 | 171 |
| KR | 32 | 32 |
| CN | 12 | 16 |
| JP | 10 | 13 |
| Etc. | 72 | 36 |
| Total | 348 | 268 |

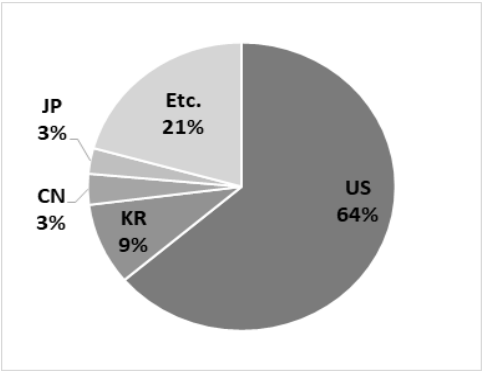


Fig. 3. Percentage of patents related to smart connectivity systems

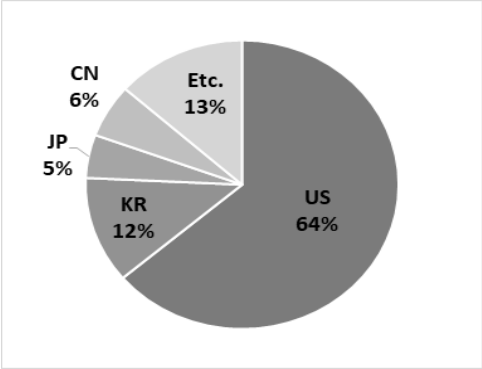


Fig. 4. Percentage of patents related to smart guide systems

## Ⅳ. Conclusion

In this study, a total number of 616 open or registered US patents from 1990 to 2017 were analyzed using a statistical text mining method. The method consisted of two steps: data preparation and clustering analysis. The patent data is segmented into two areas: smart guide systems and smart connectivity systems.

The analytical results indicate that smart guide systems and smart connectivity systems are rapidly being developed owing to the increasing urban population, which leads to greater road congestion and more accidents. Thus, governments and companies worldwide are focusing on developing appropriate technologies to mitigate these risks. Moreover, owing to developments in big data, AI, and the IoT, the means of handling these issues have changed. Current trends are moving towards zero emissions, zero accidents (i.e., automation), and zero ownership, and appropriate technologies based on ICT need to be developed to achieve these goals. The analysis indicates an increasing effort to develop smart connectivity systems and smart guide systems that use smart communication to achieve the goal of zero accidents. Thus, these systems are expected to become significantly popular in the near future.

In conclusion, it seems that the smart connectivity systems and smart guide systems are playing key roles in the development of smart mobility. Thus, establishing and developing mesh networks is essential for future technology in which communications between machines become decentralized in the shift towards the Fourth Industrial Revolution.

## REFERENCES

[1] FIT Consulting Inc, [online] https://www.fticonsulting.com/about/newsroom/press-releases/fti-consulting-projects-us-online-retail-sales-to-top-1-trillion-by-2027.

[2] United Nations, [online] https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html.

[3] Benjamin Chneider, CityLab, 2018, [online] https://www.citylab.com/transportation/2018/02/traffics-mind-boggling-economic-toll/552488/.

[4] Mark Boada, Fleet Management Weekly, [online] https://www.fleetmanagementweekly.com/qa-lukas-neckermann-fleet-mobility-revolution/.
[5] ITU, [Online] https://www.itu.int/en/ITU-T/focusgroups/ssc/Pages/default.aspx.
[6] Toru Ishida, “Digital city Kyoto”, Communication of the ACM, Vol. 45, Issue. 2, July 2002
[7] Nicos Komninso, “the architecture of intelligent cities”, 2ndinternationalconferenceonintelligentenvironment,pp.13-20,July2006.
[8] Ewelina Julita, Tomaszewska, Adrian Florea, “urban smart mobility in the scientific literature---bibliometric analysis”, international society for manufacturing service and management engineering, Vol. 10, Issue. 2, July 2018.
[9] Leif edvinsson, “aspect on the city as knowledge tool”, journal of knowledge management, Vol. 10, Issue. 5, September 2006.
[10] Chiara Garau, Francesca Masala, Francesco Pinna, “Cagliari and smart urban mobility: analysis and comparison”, Vol. 56, pp. 35-46, July 2016
[11] Andrés Camero, Enrique Alba, “Smart City and information technology: A review”, cities, vol. 93, pp. 84-94, October 2019.
[12] Rasha F. Elgazzar, Rania F. El-Gazzar, "Smart Cities, Sustainable Cities, or Both? A Critical Review and Synthesis of Success and Failure Factors," in Proceedings of the 6th International Conference on Smart Cities and Green ICT Systems (SMARTGREENS 2017), pp. 250-257, April 2017.
[13] Dietmar Harhoff, Bronwyn H. Hall, Georg von Graevenitz, Karin Hoisl, “The strategic use of patents and its implications for enterprise and competition policies”, 2014.
[14] Heidi L. Williams, “How do patents affect research investments?”, Annual Review of Economics, Vol. 9, pp. 441-469, January 2017. Beyond Patents: Assessing the value and impact of research investment.
[15] National Academy of Sciences, Beyond Patents: Assessing the value and impact of research investments, 2017.
[16] Hazel V. J. Moir, “What are the costs and benefits of patent systems?”, 2008.
[17] Ruchi Shamar, K. K. Saxena, “Strengthening the patent regime: Benefits for developing countries – a survey”, Journal of Intellectual Property Rights, Vol. 17, pp. 122-132, March 2012.
[18] S. Sathya, N. Rajendran, “A review on text mining techniques”, International Journal of Computer Science Trends and Technology (IJCST), Vol. 3, Issue 5, pp. 274-284, September-October 2015.
[19] Besim Bilalli, Alberto Abelló, Tomàs Aluja-Banet, Robert Wrembel, “Intelligent assistance for data pre-processing”, Computer Standards & Interfaces, Vol. 57, pp. 101-109, March 2018.
[20] Gautier Daras, Bruno Agard, Bernard Penz, “A spatial data pre-processing tool to improve the quality of the analysis and to reduce preparation duration”, Computers & Industrial Engineering, Vol. 119, pp. 219-232, May 2018.
[21] N. Yogapreethi, S. Maheswari, “A review on text mining in data mining”, International Journal on Soft Computing (IJSC), Vol. 7, Issue 2, pp. 1-8, August 2016.
[22] Lincoln A. Mullen, Kenneth Benoit, Os Keyes, Dmitry Selivanov, Jeffrey Arnold, “Fast, consistent tokenization of natural language text”, Journal of Open Source Software, Vol. 3, Issue 23, 655, March 2018.
[23] Tiara Shanie, Jadi Suprijadi, Zulhanif, “Text grouping in patent using adaptive k-means clustering algorithm”, AIP Conference Proceedings, Vol. 1827, Issue 1, 00241 pp. 1-9, March 2017.
[24] Kasper Welbers, Wouter Van Atteveldt, Kenneth Benoit, “Text analysis in R”, Communication Methods and Measures, Vol. 11, Issue 4, pp. 245-265, November 2017.
[25] Stefano Ferilli, Floriana Esposito, Domenico Grieco, “Automatic learning of linguistic resources for stopword removal and stemming from text”, Procedia Computer Science, Vol. 38, pp. 116-123, 2014.
[26] Dian Sa’adillah Maylawati, Wildan Budiawan Zulfikar, Cepy Slamet, Muhammad Ali Ramdhani, Yana Aditia Gerhana, “An improved of stemming algorithm for mining Indonesian text with slang on social media”, The 6th International Conference on Cyber and IT Service Management (CITSM), pp. 1-6, August 2018.
[27] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, Huan Liu, “Feature selection: A data perspective”, ACM Computing Surveys (CSUR), Vol. 50, Issue 6, January 2018.
[28] A. Fauzi, E. B. Setiawan, Z. K. A. Baizal, “Hoax news detection on Twitter using term frequency inverse document frequency and support vector machine method”, Journal of Physics: Conference Series, Vol. 1192, Issue 1, pp. 1-6, March 2019.
[29] Inbal Yahav, Onn Shehory, David Schwartz, “Comments mining with TF-IDF: The inherent bias and its removal”, IEEE Transactions on Knowledge and Data Engineering, Vol. 31, Issue 3, March 2019.
[30] Sanjiv K. Bhati, “Adaptive k-means clustering”, 2004.
[31] E. Laxmi Lydia, P. Govindsamy, S.K. Lakshmanaprabu, D. Ramya, “Document clustering based on text mining

k-means algorithm using Euclidean distance similarity", Journal of Advanced Research in Dynamical and Control Systems, Vol. 10, Issue 2, pp. 208-214, April 2018.
[32] Shizhen Zhao, Wenfeng Li, Jingjing Cao, "A user-adaptive algorithm for activity recognition based on k-means clustering, local outlier factor, and multivariate Gaussian distribution", Sensors, Vol. 18, Issue 1850, pp. 1-17, June 2018.
[33] Neha Garg, R. K. Gupta, "Clustering techniques for text mining: A review", International Journal of Engineering Research, Vol. 5, Issue 4, pp. 241-243, April 2016.
[34] A. P. Reynolds, G. Richards, V. J. Rayward-Smith, "The application of k-medoids and PAM to the clustering of rules", Lecture Notes in Computer Science, Vol. 3177, pp. 173-178, August 2004.
[35] Mahendra Tiwari, Randhir Singh, "Comparative investigation of k-means and k-medoid algorithm on iris data", International Journal of Engineering Research and Development, Vol. 4, Issue 8, November 2012.
[36] Preeti Arora, Dr. Deepali, Shipra Varshney, "Analysis of k-means and k-medoids algorithm for big data", International Conference on Information Security & Privacy, Vol. 78, pp. 507-512, December 2015.
[37] Artur Starczewski, Adam Krzyzak, "Performance evaluation of the Silhouette Index", International Conference on Artificial Intelligence and Soft Computing, Vol. 9120, pp. 49-58, June 2015.
[38] Josc Maria Luna-Romera, Maria del Mar Martinez-Ballesteros, Jorge Garcia-Gutierrez, Jose C. Riquelme-Santos, "An approach to Silhouette and Dunn clustering indices applied to big data in spark", Advances in Artificial Intelligence, pp. 160-169, September 2016.
[39] S. Govinda Rao, A. Govardhan, "Performance validation of the modified k-means clustering algorithm clusters data", International Journal of Scientific & Engineering Research, Vol. 6. Issue 10, October 2015.
[40] Zahid Ansari, M. F. Azeem, Waseem Ahmed, "Quantitative evaluation of performance and validity indices for clustering the web navigational sessions", World of Computer Science and Information Technology Journal (WCSIT), Vol. 1, Issue 5, pp. 217-226, June 2011.

## Authors

Khaled Sulaiman Khalfan Sulaiman Alkaabi received the B.S. degree in Industrial Management Engineering at Korea University, South Korea, in 2016. Currently, he is pursuing the M.S. degree in Industrial Management Engineering at Korea University, South Korea. His research interests include Text Mining, Smart Mobility, and Transportation Logistics.

Jiwon Yu received the B.S. degree in Industrial & Management Engineering from Hansung University, Korea, in 2014. received the M.S. degrees in International Logistics & Port Management from Pusan National University, Korea, in 2016. She is currently the Ph.D. in Dpt. of Industrial Management Engineering from Korea University, Korea. She is interbound in Logistics Technology Management, Text Mining, and Supply Chain Management.