R&D Perspective Social Issue Packaging using Text Analysis*

William Xiu Shun Wong** · Namgyu Kim***

Abstract

In recent years, text mining has been used to extract meaningful insights from the large volume of unstructured text data sets of various domains. As one of the most representative text mining applications, topic modeling has been widely used to extract main topics in the form of a set of keywords extracted from a large collection of documents. In general, topic modeling is performed according to the weighted frequency of words in a document corpus. However, general topic modeling cannot discover the relation between documents if the documents share only a few terms, although the documents are in fact strongly related from a particular perspective. For instance, a document about “sexual offense” and another document about “silver industry for aged persons” might not be classified into the same topic because they may not share many key terms. However, these two documents can be strongly related from the R&D perspective because some technologies, such as “RF Tag,” “CCTV,” and “Heart Rate Sensor,” are core components of both “sexual offense” and “silver industry.” Thus, in this study, we attempted to discover the differences between the results of general topic modeling and R&D perspective topic modeling. Furthermore, we package social issues from the R&D perspective and present a prototype system, which provides a package of news articles for each R&D issue. Finally, we analyze the quality of R&D perspective topic modeling and provide the results of inter- and intra-topic analysis.

Keyword: Big Data, Data Mining, Text Mining, Topic Modeling
1. Introduction

In recent years, the advances in Internet technologies and communications have influenced the manner in which people access information. With the rapid growth in the amount of information available in the electronic media, a large amount of data, which is also called “big data,” has been generated over the last few years. Big data can be utilized not only for analyzing current phenomena, but also for predicting the direction and trend of future eventualities. By utilizing big data sources, corporations can identify rapidly emerging issues, analyze trends, and enhance the efficiency and efficacy of corporate research and development (R&D) technology planning and marketing (Hong and Kim, 2014).

Big data are usually divided into two broad categories: structured and unstructured. The term structured data refers to data in a format that fits in a database and can conveniently be applied in computer programs, while unstructured data are usually better understood by humans and frequently include text and multimedia content, such as word processing files, PDF files, video, digital images, and social media posts. In this study, unstructured text data were our main concern.

Unstructured text data can be easily obtained through different social media platforms, such as news articles, blogs, online forums, and social media systems, where these unstructured text data contain both information and opinions (Kim, 2012). Therefore, big data research has become an important issue in different research fields, such as politics, economics, and business, in order to discover the unprecedented value in unstructured text data (Albright, 2006; Han et al., 2011). Big data analytics is performed to examine the large volume of unstructured text data sets, with the purpose of discovering hidden patterns, unknown correlations, market trends, customer preferences, and other useful business information (Liebowitz, 2013). High-performance analytics, such as data mining, predictive analytics, and statistical analysis, are commonly used to process the relevant data more efficiently.

Traditionally, research applying data mining investigates large masses of numeric and categorical data in order to discover valuable insights into the past, present, and future. However, the utility of data mining is limited in terms of handling huge amounts of unstructured textual documents. Therefore, a relatively new technique that is emerging from data mining, text mining, has been proposed to perform knowledge discovery from an unstructured text corpus (Weiss et al., 2012). Text mining is usually applied to extract valuable information from natural language text. In summary, the steps of text mining are as follows: (1) Convert the input text into a structured format that can be stored in a database, (2) construct various models using data mining techniques, and (3) interpret the outcome and discover new insights (Tseng et al., 2007; Wang and Ohsawa, 2012).

An important text-mining application is topic modeling, used mainly to discover a set of “topics” that are contained in a large collection of documents (Blei et al., 2003; Hofmann, 2001). Topic modeling describes a topic as a recurring pattern of co-occurring words. Measures such as term frequency-inverse document frequency (TF-IDF) are usually applied to measure the relative importance of terms in each document (Han, 2011; Provost and Fawcett, 2013). In most cases, topic modeling allows each document to be related to
multiple topics. This desirable characteristic has led to its popularity in areas such as national issue excavation through news articles and social issue extraction from tweets to facilitate election and marketing decision making (Bae et al., 2013; Hyun et al., 2013).

In general, topic modeling is performed according to the weighted frequency of words in a document corpus. For example, if some words appear simultaneously in various documents, the documents can be considered similar and grouped into the same topic. However, by using general topic modeling, the relation between documents cannot be recognized if the documents share only a few terms, although they are in fact strongly related from a particular perspective. For instance, a document about “sexual offense” and another document about “silver industry for aged persons” may not be classified into the same topic, because they may not share many key terms. However, these two documents can be strongly related from an R&D perspective because some technologies, such as “RF Tag,” “CCTV,” and “Heart Rate Sensor,” are core components of both “sexual offense” and “silver industry.” <Figure 1> further explains this notion.

<Figure 1> shows an explanatory example of the R&D perspective reorganization of news articles. The left hand side of the diagram shows the news articles arranged according to their subject categories. If general topic modeling is applied to this document corpus, the results reflect mainly the general key terms in this corpus. In the results, for instance, there is a strong possibility that the news articles in the “Sports” category will contain co-occurrence terms such as “football,” “baseball,” and “exercise.” However, if R&D perspective topic modeling is applied to this document corpus, the news articles that share many R&D key terms can be grouped, as

---

**Sports**
- Fantasy Football Ranking 2015
- Student Charged After **Drone** Crash During Kentucky Football Game
- NC Rallies Late to Beat Doosan

**IT Science**
- Museum Specimens Find New Life Online
- Animals Spy a New Enemy: **Drones**
- LG Makes Big Push on IoT

**Lifestyle**
- Korea Dreams of **Drone** Delivery
- 36 Hours in Eastern Tokyo
- Asian Spicy Curry Pops Up in Chinatown

**World**
- Turkey Confirms **Drone** Was Shot Down
- **Drone** Captures Indonesia’s Forest Fires
- Hong Kong Bookstores Display Beijing’s Clout

---

**R&D Perspective Index**
- Unmanned Aerial Vehicle; **Drone**
  - Student Charged After **Drone** Crash During Kentucky Football Game
  - Animals Spy a New Enemy: **Drones**
  - Korea Dreams of **Drone** Delivery
  - Turkey Confirms **Drone** Was Shot Down
  - **Drone** Captures Indonesia’s Forest Fires

---

<Figure 1> Notion of R&D Perspective Document Reorganization
shown on the right hand side of the diagram. The figure shows that five news articles from four different categories, “Sports,” “IT Science,” “Lifestyle,” and “World,” can compose a new single-issue group, “Unmanned Aerial Vehicle.”

Thus, in this study, we attempted to discover the differences between the results of general and R&D perspective topic modeling. Furthermore, we package social issues from the R&D perspective. Finally, we present snapshots of a prototype system that provides a package of news articles for each R&D issue. This paper is organized as follows. First, as an introductory statement, the general background of text analysis, text mining, clustering analysis, and topic modeling is presented in Section 2. Next, the overall process of the proposed methodology, the results of related experiments, and a prototype system are described in Section 3. The performance evaluations of our methodology are presented in Section 4. Finally, Section 5 concludes this paper, describing the contributions, the study limitations, and the key directions for the future research.

2. Related Work

2.1 Text Mining

Text appears in various forms and is widely used for exchanging information in the real world (Hearst, 1999). Text mining is a frequently used method for discovering the valuable knowledge contained in natural language text. Traditional data mining attempts to seek interesting patterns in large structured databases, while text mining utilizes the advantages of data mining techniques for discovering interesting information in the unstructured textual data of different kinds of documents, such as news articles, reports, patent information, and product descriptions (Holton, 2009; Hyun et al., 2013). Further, text mining itself is also a new research area, which draws on general statistics, information retrieval, machine learning, artificial intelligence, natural language processing, and computational linguistics (Feldman and Sanger, 2007; Provost and Fawcett, 2013).

In detail, text mining transposes the words and phrases in unstructured data into structured vectors (Stanvrianou, 2007). By using text analytics software and analysis models, the generated indices can be stored in a structured database and incorporated in other analyses, such as predictive data mining projects and unsupervised learning methods. In addition, researchers and industrial experts can use text analytics to gain insight into content-specific values, such as sentiment, emotion, intensity, and relevance (Nagaraj et al., 2014). Different techniques are utilized in the text mining area, such as text categorization, text clustering, entity extraction, sentiment analysis, document summary, and entity-relation modeling (Agrawal and Batra, 2013; Mooney et al., 2006). In the study of Meyer et al. (2008), text mining was proposed as a method for preprocessing documents. Moreover, text mining is also actively applied in identifying technology trends for R&D planning (Wang et al., 2010).

Usually, text data can be obtained from various media sources. By analyzing the media content, it is possible to obtain strategic insight and intelligence on media sources and issues reported in the media (Macnamara, 2005). Mass media, as one of the world’s largest databases, reflects the opinions and perceptions of the
general public through its reporting of current issues and trends. Currently, most of the issues and trends are published as news articles provided by online news services. Additionally, news articles deliver professional information in various areas. For instance, there is a large volume of news articles related to advanced technologies and skills. These news texts contain a wide range of information related to the concepts, markets, and social impact of new technologies. Therefore, the trends of technologies and the current status of product innovation activities can be identified by analyzing these news articles.

News articles constitute a source of unstructured text data, and therefore, the capability of metadata search to handle different data formats needs to be considered. Further, technology for de-duplicating information, eliminating noise, and enhancing the accuracy of feature extraction is also required for news texts analysis. Many attempts to achieve news analysis using text mining have been reported. For instance, Yoon (2012) investigated the weak signals for long-term business opportunities from Web news articles, while Hong and Kim (2014) identified emerging technologies from an immense volume of news texts.

2.2 Document Clustering and Topic Modeling

Clustering is an unsupervised learning method that reduces the amount of information to be analyzed by grouping similar data items together (Barbakh et al., 2009; Jain et al., 1999; Romero et al., 2008). Document clustering is a text mining application that finds the clusters of documents that are similar to each other within a group and dissimilar to documents in other groups. In other words, the objective of document clustering is to maximize the similarity among documents within the same clusters and to maximize the dissimilarity of documents among documents in different clusters. Document clustering is a fundamental research issue that is related to many other applications, such as document organization, browsing, summarization, and classification (Aggarwal and Zhai, 2012; Cai et al., 2011; Lu et al., 2011; Ng et al., 2002; Xu and Gong, 2004; Xu et al., 2003).

Each document can belong to only one cluster when hard clustering techniques are used, while each document can be included in multiple clusters when soft clustering techniques are used.

Currently, topic modeling is widely used as a special implementation of document clustering. Document clustering and topic modeling are two closely related tasks that can mutually benefit each other. In topic modeling, documents are grouped into clusters as in document clustering. However, topic modeling assigns a subject to each cluster by listing the core terms of the cluster, while general document clustering does not assign a meaningful name to each cluster. In a study in the literature, Lu et al. (2011) investigated the clustering performance of probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA). They used LDA and PLSA to model the corpus and treated each topic as a cluster. Topic models are probabilistic generative models initially created to model texts and identify the latent semantics underlying a document corpus. Topic modeling can identify multiple latent semantic topics in a document corpus, where each topic is represented as a multinomial distribution over a given vocabulary and each document is a mixture of
hidden topics (Xie and Xing, 2013).

In topic modeling, the well-known TF-IDF formulation has been used for weighting terms. Specifically, TF-IDF computes term frequency with inverse document frequency. Term frequency is the number of times a term occurs in a document, while inverse document frequency reflects whether a term is common or rare across the document corpus. Inverse document frequency methods check the number of documents containing the term and reverse the scaling (Aizawa, 2003; Carnerud, 2014). This calculation indicates the importance of a word in a particular document in a document corpus, because TF-IDF is able to capture the amount of information a term provides. A term that is relatively rare and appears frequently in only a few documents receives a high value.

With regard to real-world applications, topic models have been applied to newspaper corpora in order to discover social issues and trends over time. For instance, topic models were used to analyze a collection of 18th century American newspapers to reveal important issues (Newman and Block, 2006). Griffiths and Steyvers (2004) identified research topic trends by applying topic models to abstracts of scientific papers, while Sun (2014) applied a topic modeling technique to identify scientific topics in a collection of American newspapers. In addition, topic modeling has been utilized in many other areas, such as natural language processing, the analysis of online shopping mall product reviews, crime prediction, and repository establishment, through text categorization (Provost and Fawcett, 2013; Myung et al., 2008; Liu, 2012; Fan et al., 2006; Sebastiani, 2006). Furthermore, based on R&D keywords, the social issues were extracted from the Korean newspaper using topic modeling and reorganized by using the clustering technique (Wong and Kim, 2015), while Kim et al. (2014) utilized the topic modeling technique and social network analysis to group the issues from the user perspective.
3. Proposed Methodology

3.1 Research Scope

In this section, the methodology of the proposed model is described with some examples. <Figure 2> shows an overview of the entire process.

The four core processes of this methodology are as follows. First, we collect a large volume of news articles published in a certain period of time by Web crawling (1). The general topic analysis module (2) performs traditional topic modeling and derives the social issues on the basis of document similarity. In the R&D perspective topic analysis stage (3), in contrast, we perform another topic modeling, emphasizing R&D terms and neglecting non-R&D terms. Clearly, the results of process (3) are significantly different from the results of process (2). To reveal the difference more clearly, the topic mapping stage (4) maps social issues from general topic modeling to R&D topics from R&D perspective topic modeling.

One of the main purposes of this study was to reorganize the documents from an R&D perspective by using topic modeling; therefore, we needed to construct an R&D lexicon from R&D-related documents. <Figure 3> shows the R&D lexicon construction process. First, we obtain the patent documents from the Intellectual Property Rights (IPR) information services. After parsing the documents, we extract frequent terms and place them in a temporary list. However, the list may contain a number of noisy terms and non-R&D terms. Therefore, we need to perform the next step: stop word removal. First, we remove pronouns, prepositions, and conjunctions. Even after removing the stop words, the list may still contain meaningful but non-R&D terms. Thus, general terms that frequently appear also in general documents, such as news articles, should be removed from the list. After four domain experts have refined the temporary list, we finally obtain an R&D lexicon. The refined R&D lexicon is used as a start list in R&D perspective topic analysis and evaluation processes.

3.2 Topic Analysis and Topic Mapping

As shown in <Figure 2>, the topic analysis process is divided into two parts: general topic analysis and R&D perspective topic analysis. In general, topic analysis is performed based on the TF-IDF measure shown as:

\[
TF-IDF(d, t) = TF(d, t) \times IDF(t)
\]

\[
TF(d, t) = \begin{cases} 
0 & \text{if } freq(d, t) = 0 \\
\frac{1 + \log(1 + \log(freq(d, t)))}{\log|d|} & \text{otherwise}
\end{cases}
\]

\[
IDF(t) = \log \left( \frac{|d|}{|d_t|} \right)
\]
It is used to measure the relative importance of each word in each document. Because topic analysis itself is not our main concern and has been introduced in many previous studies, we do not describe it in detail in this paper.

The net results of this process are represented as a document/topic correspondence matrix. We can calculate the document topic weight of an individual document toward each topic. The document topic weight of a document toward a topic is the normalized sum of the TF-IDF weightings for each term in the document multiplied by the topic weights. We can convert the value of the document topic weight into a binary value using a certain threshold. The threshold is set to the mean plus one standard deviation of all document topic weights for the corresponding topic. In the binary form of the document/topic correspondence matrix, every value in each cell shows whether the document can be included in a particular topic or not. <Figure 4> shows an example of the document/topic correspondence matrix. In this figure, for instance, it can be seen that document “Doc_2” is related to the topics “Topic_1” and “Topic_4.”

| Topic_1 | Topic_2 | Topic_3 | Topic_4 | Topic_5 |
|---------|---------|---------|---------|---------|
| Doc_1   | 0       | 0       | 0       | 0       | 1       |
| Doc_2   | 1       | 0       | 0       | 1       | 0       |
| Doc_3   | 1       | 0       | 0       | 1       | 0       |
| Doc_4   | 0       | 1       | 1       | 0       | 0       |
| Doc_5   | 0       | 1       | 1       | 0       | 0       |
| Doc_6   | 1       | 0       | 0       | 1       | 1       |
| Doc_7   | 1       | 0       | 0       | 0       | 1       |
| Doc_8   | 0       | 1       | 0       | 0       | 0       |

<Figure 4> Example of Document/Topic Correspondence Matrix

The main difference between general topic analysis and R&D perspective topic analysis is the application of the R&D lexicon. In the process of R&D perspective topic analysis, we can acquire a second document/topic correspondence matrix that is significantly different from the general topic analysis matrix. By utilizing the R&D lexicon as a start list of topic analysis, we can emphasize only R&D terms and neglect non-R&D terms. Using the two different versions of topic analysis, finally, we can reveal the relationship between each R&D topic and various social issues. The packaging of social issues from the R&D perspective was the main objective of our study and could be accomplished by topic mapping.

The example in <Figure 5> explains the topic mapping process. <Figure 5(a)> shows the correspondence matrix of documents and social issues, while <Figure 5(b)> displays the correspondence matrix of documents and R&D topics. If a particular document is related to a particular topic, the value of the corresponding cell is marked “1.” For example, it can be seen that “Doc_1” is related to the social issues “Gen_1” and “Gen2” in <Figure 6(a)> as well as to the R&D topic “R&D_1.” By performing matrix multiplication for the two matrices, we can derive the correspondence matrix between R&D topics and social issues, as shown in <Figure 5(c)>.

In <Figure 5(c)>, the value of each cell represents the number of common documents between each social issue and each R&D topic. For example, the degree of correspondence between “Gen_1” and “R&D_1” is revealed to be “2,” because both of the two topics have “Doc_1” and “Doc_4” as their member documents. Formally, by a matrix multiplication, the A-th row in “Matrix 1” and D-th column in “Matrix 2” can produce the value of Cell(A, D) in the resultant matrix. The correspondence between the A-th row in the Matrix 1 and D-th column in Matrix 2 can be derived using the formulation as below:
R&D Perspective Social Issue Packaging using Text Analysis

79

3.3 Prototype—R&D Perspective Social Issue Packaging

In this section, we demonstrate our service scenario by providing a simple prototype. We first explain the data used in this research. Then, the results of topic mapping are described in detail. Finally, we present a prototype that provides a package of social issues and individual news articles from the R&D perspective.

3.3.1 Data Description

In this subsection, we introduce the real-world data used in our research. As shown in Figure 6, 3,000 articles published between July 2012 and June 2013 were randomly selected from each of 8 different sections of the South Korean news portal site “N.” The sections were “IT Science,” “Politics,” “Economics,” “Community,” “Lifestyle,” “World,” “Sport,” and “Entertainment.” In our opinion, the news portal site “N” was the most appropriate source for our experiment, because it gathers and re-publishes news articles from multiple representative news media outlets in South Korea. This implies that the site “N” does not represent the biased opinion of a specific group. Thus, a total of 24,000 news articles from the site “N” were selected and used as the input of the topic modeling process.

\[
\text{Corr}(A,D) = \sum_{i=1}^{n} (\text{Mtrx 1. Cell}(A,i) \times \text{Mtrx 2. Cell}(i,D))
\]

where “n” represents the number of documents used. The larger the number of common documents between an R&D topic and a social issue in a certain combination, the higher the degree of correspondence between the two topics. Thus, we can discover the social issues related to a particular R&D topic on the basis of the degree of their correspondence.

3.3 Prototype—R&D Perspective Social Issue Packaging

In this section, we demonstrate our service scenario by providing a simple prototype. We first explain the data used in this research. Then, the results of topic mapping are described in detail. Finally, we present a prototype that provides a package of social issues and individual news articles from the R&D perspective.

3.3.1 Data Description

In this subsection, we introduce the real-world data used in our research. As shown in Figure 6, 3,000 articles published between July 2012 and June 2013 were randomly selected from each of 8 different sections of the South Korean news portal site “N.” The sections were “IT Science,” “Politics,” “Economics,” “Community,” “Lifestyle,” “World,” “Sport,” and “Entertainment.” In our opinion, the news portal site “N” was the most appropriate source for our experiment, because it gathers and re-publishes news articles from multiple representative news media outlets in South Korea. This implies that the site “N” does not represent the biased opinion of a specific group. Thus, a total of 24,000 news articles from the site “N” were selected and used as the input of the topic modeling process.

\[
\text{Corr}(A,D) = \sum_{i=1}^{n} (\text{Mtrx 1. Cell}(A,i) \times \text{Mtrx 2. Cell}(i,D))
\]

where “n” represents the number of documents used. The larger the number of common documents between an R&D topic and a social issue in a certain combination, the higher the degree of correspondence between the two topics. Thus, we can discover the social issues related to a particular R&D topic on the basis of the degree of their correspondence.

3.3 Prototype—R&D Perspective Social Issue Packaging

In this section, we demonstrate our service scenario by providing a simple prototype. We first explain the data used in this research. Then, the results of topic mapping are described in detail. Finally, we present a prototype that provides a package of social issues and individual news articles from the R&D perspective.

3.3.1 Data Description

In this subsection, we introduce the real-world data used in our research. As shown in Figure 6, 3,000 articles published between July 2012 and June 2013 were randomly selected from each of 8 different sections of the South Korean news portal site “N.” The sections were “IT Science,” “Politics,” “Economics,” “Community,” “Lifestyle,” “World,” “Sport,” and “Entertainment.” In our opinion, the news portal site “N” was the most appropriate source for our experiment, because it gathers and re-publishes news articles from multiple representative news media outlets in South Korea. This implies that the site “N” does not represent the biased opinion of a specific group. Thus, a total of 24,000 news articles from the site “N” were selected and used as the input of the topic modeling process.

\[
\text{Corr}(A,D) = \sum_{i=1}^{n} (\text{Mtrx 1. Cell}(A,i) \times \text{Mtrx 2. Cell}(i,D))
\]

where “n” represents the number of documents used. The larger the number of common documents between an R&D topic and a social issue in a certain combination, the higher the degree of correspondence between the two topics. Thus, we can discover the social issues related to a particular R&D topic on the basis of the degree of their correspondence.

3.3 Prototype—R&D Perspective Social Issue Packaging

In this section, we demonstrate our service scenario by providing a simple prototype. We first explain the data used in this research. Then, the results of topic mapping are described in detail. Finally, we present a prototype that provides a package of social issues and individual news articles from the R&D perspective.

3.3.1 Data Description

In this subsection, we introduce the real-world data used in our research. As shown in Figure 6, 3,000 articles published between July 2012 and June 2013 were randomly selected from each of 8 different sections of the South Korean news portal site “N.” The sections were “IT Science,” “Politics,” “Economics,” “Community,” “Lifestyle,” “World,” “Sport,” and “Entertainment.” In our opinion, the news portal site “N” was the most appropriate source for our experiment, because it gathers and re-publishes news articles from multiple representative news media outlets in South Korea. This implies that the site “N” does not represent the biased opinion of a specific group. Thus, a total of 24,000 news articles from the site “N” were selected and used as the input of the topic modeling process.

\[
\text{Corr}(A,D) = \sum_{i=1}^{n} (\text{Mtrx 1. Cell}(A,i) \times \text{Mtrx 2. Cell}(i,D))
\]

where “n” represents the number of documents used. The larger the number of common documents between an R&D topic and a social issue in a certain combination, the higher the degree of correspondence between the two topics. Thus, we can discover the social issues related to a particular R&D topic on the basis of the degree of their correspondence.

3.3 Prototype—R&D Perspective Social Issue Packaging

In this section, we demonstrate our service scenario by providing a simple prototype. We first explain the data used in this research. Then, the results of topic mapping are described in detail. Finally, we present a prototype that provides a package of social issues and individual news articles from the R&D perspective.

3.3.1 Data Description

In this subsection, we introduce the real-world data used in our research. As shown in Figure 6, 3,000 articles published between July 2012 and June 2013 were randomly selected from each of 8 different sections of the South Korean news portal site “N.” The sections were “IT Science,” “Politics,” “Economics,” “Community,” “Lifestyle,” “World,” “Sport,” and “Entertainment.” In our opinion, the news portal site “N” was the most appropriate source for our experiment, because it gathers and re-publishes news articles from multiple representative news media outlets in South Korea. This implies that the site “N” does not represent the biased opinion of a specific group. Thus, a total of 24,000 news articles from the site “N” were selected and used as the input of the topic modeling process.
In order to construct the R&D lexicon, we used the summary section of patent documents. A total of 10,000 cases of the “Daily Necessity,” “Electricity,” and “Mechanical Engineering” categories between July 2011 and March 2013 were obtained from WIPS, an online worldwide patent information service provider in South Korea. According to the “R&D Lexicon Construction” process described in <Figure 3>, a total of 15,043 terms were included in the final R&D lexicon.

3.3.2 Results of Topic Mapping (Medical and Transportation Cases)

In the general topic analysis and R&D perspective topic analysis stage, we used the commercial data-mining tool “SAS Enterprise Miner” to derive social issues and R&D topics, respectively. A total of 100 topics were generated through a general topic analysis of the 24,000 news articles. The main difference between the general topic analysis and the R&D perspective topic analysis is that the latter derives topics considering only terms contained in the R&D lexicon generated by the process shown in <Figure 3>. By performing matrix multiplication for the two types of topic/document correspondence matrix, we can determine the relevance between R&D topics and social issues, as depicted in <Figure 5>.

Each of the 100 R&D topics has its relevant social issues. Instead of listing all the results, in this subsection we describe the process for two selected target R&D domains. We identified the representative R&D topics in the selected domain, and then, discovered the relevant social issues of the representative R&D topics. For this purpose, we selected “Medical” and “Transportation” as the target R&D domains, because there has been a considerable amount of investment and research in these two fields. The detailed process for identifying the representative R&D topics in each selected R&D domain is shown in <Figure 7>.
Criteria documents for a specific R&D domain were selected and used as references for the filtering process. We used patent documents and technical reports submitted to the government of the Republic of Korea as criteria documents for the medical and transportation domains. In this subsection, we explain the medical domain case as an example. By filtering, R&D terms that occur in both the medical domain criteria documents and the R&D lexicon were extracted. This list of medical R&D terms was used to measure the relevance of each news article to the subject of medical R&D. For this purpose, the frequency analysis process counts the frequency of the medical domain–related R&D terms in each news article. The R&D topic/domain correspondence analysis is aimed to find the representative R&D topics in the medical domain on the basis of the frequency of medical R&D terms in each article. In this process, the frequency of medical R&D terms in each article is summarized for each R&D topic. Then, we calculated the average frequency of medical R&D terms per article for each topic to find the topics that are strongly related to medical R&D.

After identifying the representative medical R&D topics, we selected the top three topics for the two R&D domains, as shown in <Figure 8>. For the medical domain, 148 medical R&D terms, including adenovirus, cell strain, bio, and vaccine, were used. For the transportation domain, 40 transportation R&D terms exist, such as traffic, vehicle, communication, and roadside. Furthermore, the related information of the top three R&D topics, such as topic id, topic keywords, term frequency, document frequency, and the average of term frequency, are also shown in <Figure 8>.

In <Figure 8>, it can be seen that the best matching topic of medical R&D is Topic 74 with the keywords “patient, virus, treatment, infection, disease” while the best matching topic of transportation R&D is Topic 64 with keywords “vehicle, parking lot, mobile, fire disaster, program.” Given a certain R&D topic, as explained in <Figure 5> in Section 3.2, we can identify social issues that are related to it.

According to the topic mapping, which is the correspondence matrix between R&D topics and social issues, the social issues that correspond to medical and transportation R&D were extracted.

| Medical Domain (148 terms including adenovirus, cell strain, bio, vaccine,...) |
|-------------------------------|-----------------|-----------|-----------------|
| Topic ID | Topic Keywords | Term Freq. | Doc Freq. | Average Term Freq. |
|---------|----------------|------------|-----------|--------------------|
| 74      | patient, virus, treatment, infection, disease | 3756 | 1070 | 3.51028 |
| 54      | medical, Jinju, patient, health, treatment | 2575 | 887 | 2.90304 |
| 6       | health, treatment, patient, disease, weight | 4809 | 1673 | 2.87448 |

| Transportation Domain (40 terms including traffic, vehicle, communication, roadside,...) |
|---------------------------------------------|---------------|-----------|-----------------|
| Topic ID | Topic Keywords | Term Freq. | Doc Freq. | Average Term Freq. |
|---------|----------------|------------|-----------|--------------------|
| 64      | vehicle, parking lot, mobile, fire disaster, program | 2801 | 1151 | 2.43354 |
| 5       | engine, fuel consumption ratio, vehicle, drive, design | 3232 | 1355 | 2.38524 |
| 22      | road, vehicle, driver, drive, transportation | 4013 | 1787 | 2.24566 |

<Figure 8> Top Three R&D Topics in Medical and Transportation Domains
As part of the topic mapping results, the best matching R&D topics of the selected domains and their corresponding social issues were identified as shown in <Figure 9>.

### 3.3.3 Prototype of R&D Perspective Issue and News Packaging Service

In this subsection, we present a simple prototype of the proposed methodology for packaging social issues and individual news items from the R&D perspective. In this stage, we focus mainly on potential service scenarios and thus present some screenshots of the main user interfaces. <Figure 10> shows the start page of the R&D perspective issue and news packaging service. This page contains conventional sections, such as a category menu bar and a list of individual articles. The original contribution is located in the right hand section of <Figure 9>. The new function “R&D Perspective Index” provides a list of R&D topics with the percentage of the amount of articles related to the corresponding R&D topic. From this start page, a user can navigate to specific R&D-related pages by clicking an R&D topic in the right section.

Let us suppose a situation where a user wishes to investigate issues and individual articles related to the R&D topic “Drone.” By clicking the R&D topic “Drone” in the “R&D Perspective Index” section, the user can investigate the R&D topic “Drone” in detail. First, many articles from various categories could be related to “Drone.” The category distribution of articles about “Drone” may constitute useful information for perceiving the practical usage of “Drone.” Our prototype presents the category distribution of articles about a specific R&D topic in the form of a pie chart, shown in <Figure 11>. In the category table section, the user can check the number of news articles related to “Drone” for each category. The R&D keywords for “Drone,” such as “Drone,” “UAV,” “Unmanned Aerial Vehicle,” “Aircraft,” and “Robot Plane,” that were obtained from the topic modeling process are displayed in the upper section of <Figure 11>.

| R&D Domain   | Keywords for R&D Topics                      | Keywords for Social Issues                                                                 |
|--------------|----------------------------------------------|------------------------------------------------------------------------------------------|
| Medical      | patient, virus, treatment, infection, disease | police, suspicion, police station, incident, crime, patient, treatment, hospital, symptom, disease |
|              |                                              | research, research team, professor, result, Ph.D, food, health, exercise, research, vitamin |
|              |                                              | terror, Boston, suspect, marathon, USA                                                     |
|              |                                              | Busan, clinic, China, brand, female                                                       |
| Transportation | vehicle, parking lot, mobile, fire disaster, program | model, engine, vehicle, fuel consumption ratio, automobile, typhoon, Bolaven, gale, Jeju |
|              |                                              | police, suspicion, police station, incident, crime, temperature, province, middle area, weather, Korea Meteorological Administration |
|              |                                              | accident, vehicle, driver, drive, police                                                 |
(Figure 10) Start Page of the R&D Perspective Issue and News Packaging Service

(Figure 11) Category Distribution of Articles about “Drone”
<Figure 12> Individual News Articles Related to "Drone" in "Sports" Category

<Figure 13> Individual News Articles Related to "Drone" and a Specific Social Issue
If a user wishes to investigate individual articles about “Drone” only in the “Sports” category, he/she can simply click “Sports” in the category table section. The news articles related to “Drone” in the “Sports” category are then displayed, as shown in Figure 12.

Figure 13 shows one of the most important parts of our methodology. As we provided the R&D topics and their corresponding social issues, we can list the news articles that are related to a specific combination of the R&D topic and the social issue. For example, social issues related to unmanned aerial vehicle are displayed in the right part of Figure 13. When the user clicks on the social issue “Student, Drone, Charged, Football, Unauthorized,” the related news titles are displayed in the left part of Figure 13. This structure allows users to see which social issues are highly related to a certain R&D topic. By examining the titles and content of the articles, furthermore, a user can comprehend which events and opinions have actually appeared related to the combination of R&D topic and social issue.

4. Performance Evaluation

4.1 Evaluation Process

In this section, we describe the results of the performance evaluation of our methodology. We first explain the process of preparing the data needed for the evaluation analysis. Then, as the evaluation main process is divided into two different analyses, inter- and intra-cluster, the detailed process and the results of these two analyses are presented. The performance evaluation is important, because it allows us to measure the actual performance of our proposed methodology, as well as to compare it with that of the traditional approach.

Figure 14 shows the evaluation process of
our methodology. First, we selected the criteria documents from the R&D-related document corpus as the target data. The reason for this processing step is that, because the R&D lexicon contains R&D terms from different R&D fields, if it were not performed the evaluation results would be complicated and the differences between our proposed methodology and a traditional approach would be difficult to determine. To provide simple and clear comparison results, we chose two R&D fields as the main R&D domains, the medical and transportation fields. The criteria documents were selected from these two main domains. Then, we derived the R&D terms from the criteria reports through the filtering process by using the R&D lexicon as the start list. Next, the derived R&D terms were used in the frequency analysis to obtain the frequency of the selected R&D domain terms that occurred in the correspondence matrices of document/social issue and document/R&D topic. The correspondence matrix of document/social issue is the outcome of the general topic modeling, while the correspondence matrix of document/R&D topic is the outcome of the R&D perspective topic modeling, as shown in <Figure 2>. The resulting frequency of the selected R&D domain terms in each article for each topic/issue was used in the inter- and intra-topic evaluation analysis. For the inter-topic evaluation analysis, we obtained the ratio of the valid documents to all documents, called the valid ratio. Then, we compared the distribution of the valid ratio for R&D topic modeling and general topic modeling. For the intra-topic evaluation analysis, we derived the ratio of R&D terms to all documents, called the term ratio. We also compared the distribution of the term ratio for both different types of topic modeling.

4.2 Inter-Topic Evaluation

The purpose of performing the inter-topic evaluation was to determine the topic distribution between the R&D topic modeling and general topic modeling. Therefore, we first needed to obtain the valid ratio of each topic for the correspondence matrix of either R&D topic modeling or general topic modeling. Let \( V_j \) denote a valid document that contains a number of R&D terms above a predefined threshold and \( d_j \) denote a particular document among the 24,000 documents. Then, we obtained the frequency of the valid documents in each topic, as described in (1). Next, we derived the ratio of the valid documents to all documents \( N \) for each topic, as shown in (2). The ratio derived in (2) is the valid ratio.

\[
V_j = \begin{cases} 
0 & \text{if } \text{term_cnt}(d_j) < \text{threshold} \\
1 & \text{otherwise} 
\end{cases}
\]

\[
\text{Valid_Cht}(T_i) = \frac{\sum_{j=1}^{M} V_j}{\sum_{j=1}^{M} d_j} 
\]

for each valid document \( V_j \) in topic \( T_i \) (1)

\[
\text{AVG(Valid_Cht}(T_i)) = \frac{\sum_{i=1}^{N} \text{Valid_Cht}(T_i)}{N} 
\]

Then, we arranged the valid ratios of each topic in descending order, so that the most relevant documents to the non-relevant documents could be examined in order of their ranking from top to bottom. Further, based on the descending order of the valid ratios, we derived the cumulative mean and cumulative standard deviation for the topics, in order to determine the distribution of the topics and the differences between our approach and a traditional approach.
Figure 15 Inter-Topic Evaluation (Medical Domain)
Figure 16: Inter-Topic Evaluation (Transportation Domain)

Cumulative Mean Distribution Graph (Transportation Domain)

Cumulative Mean

Number of Topics

- R&D Topic Modeling
- General Topic Modeling

Cumulative Standard Deviation Distribution Graph (Transportation Domain)

Cumulative Standard Deviation

Number of Topics

- R&D Topic Modeling
- General Topic Modeling

(Figure 16) Inter-Topic Evaluation (Transportation Domain)
<Figure 15> shows the distribution graph of the cumulative mean and cumulative standard deviation for the medical domain. In <Figure 15>, our approach, which is the R&D topic modeling, shows a steep slope as compared to the general topic modeling in the cumulative mean distribution graph. In other words, the most relevant documents of the medical domain are grouped together in the top ranking of topics, while the non-relevant documents are grouped together in the lower ranking of topics. However, for general topic modeling, the relevant documents of the medical domain are distributed almost equally into each topic, and therefore, it is difficult to determine which topic is relevant to the medical domain, resulting in a gradual slope in the graph.

In the cumulative standard deviation distribution graph, the grey area shows the saturation part of the process. If the saturation part is ignored, R&D topic modeling shows a better performance as compared to general topic modeling. This is because the higher the standard deviation of the topics, the higher the discriminative power of the topics. An examination of the end part of the curves reveals that the standard deviation of R&D topic modeling is 0.294, while the standard deviation of traditional topic modeling is 0.237. In addition, the gradient of R&D topic modeling from high to low indicates a more desirable result than does that of general topic modeling.

<Figure 16> shows the results of the inter-topic evaluation for transportation. Although the standard deviation of R&D topic modeling is higher than the standard deviation of general topic modeling, it is difficult to determine the pattern of higher ranking versus that of lower ranking. A reason for this is may be that the result varies according to different R&D fields. Another reason may be the limitation of our approach as the R&D terms extracted from the transportation criteria report were not sufficient to show a significant result. Thus, the inter-topic evaluation results of the medical domain are more reliable than those of the transportation domain.

4.3 Intra-Topic Evaluation

In this section, the evaluation of the two different topic modeling, R&D topic modeling and general topic modeling, for the two R&D domains is described. First, the frequency of the R&D terms for each document in each topic was calculated and the ratio of the R&D terms to all documents for each topic was derived, as shown in (3). The ratios are called term ratios. Further, we also calculated the standard deviation of each topic by using the frequency of R&D terms for each document in each topic according to (4). Finally, using (5), we derived the average of the standard deviation for both different topic modelings.

In intra-topic evaluation, the lower the average of the standard deviation, the better.

\[
AVG_{\text{Term\_Cnt}}(T_i) = \frac{\sum_{j=1}^{M} \text{Term\_Cnt}(d_j)}{M}
\]

for each document \(d_j\) in topic \(T_i\) \hspace{1cm} (3)

\[
STD_{\text{Term\_Cnt}}(T_i) = \sqrt{\frac{1}{M} \sum_{j=1}^{M} \left( \text{Term\_Cnt}(d_j) - AVG_{\text{Term\_Cnt}}(T_i) \right)^2}
\]

for each document \(d_j\) in topic \(T_i\) \hspace{1cm} (4)

\[
AVG_{STD\_\text{Term\_Cnt}} = \frac{\sum_{j=1}^{N} STD_{\text{Term\_Cnt}}(T_i)}{N}
\]
Figure 17: Intra-Topic Evaluation

Medical Domain

Transportation Domain

Cumulative Standard Deviation

Number of Topics

R&D Topic Modeling

General Topic Modeling

(Figure 17) Intra-Topic Evaluation
the topic modeling performance. For the medical domain, the average of standard deviation for R&D topic modeling is 0.517 and for general topic modeling 0.622. In addition, for transportation, the average of standard deviation for R&D topic modeling is 0.457 and for general topic modeling 0.471. Our approach shows a better performance with a low average value as compared to general topic modeling.

However, as our objective was to determine which topic shows the better performance results between the two different topic modelings, we decided to illustrate the results in a graph. In order to illustrate the distribution of topics, the standard deviations of each topic were sorted according to the descending order of the term ratio. This is because the high term ratio refers to more R&D terms in that topic, and thus, the more relevant the topic is to an R&D domain. The graphs were plotted according to the sorted standard deviation value; the results are shown in Figure 17.

In the case of the medical domain, the gap between the two topic modelings is wider in the high relevance topic ranking, which shows that our approach outperforms others. Although the gap becomes smaller across the low relevance topic ranking, it still shows a better result as compared to general topic modeling. For the transportation domain, the gap between the two topic modelings is too small and a significant difference between R&D topic modeling and general topic modeling is not seen in the gradient. Therefore, it is difficult to extract meaningful insight from the transportation domain results.

5. Conclusion

In this study, we examined techniques for utilizing unstructured text data. The aim of this study was to develop a method to package a social issue from the R&D perspective by using text analysis. We focused mainly on using topic modeling, a text mining application, as the main analysis method. In general, topic modeling uses the weighted frequency of words in a document corpus to discover topics. Based on our proposed methodology, the reorganization of social issues can be more effective and its results can serve as a package of news articles for each R&D issue.

Furthermore, this paper provided a prototype system based on our methodology. We also presented a comparative performance evaluation of R&D perspective topic modeling and general topic modeling, which can lead to different outcomes. The R&D perspective topic modeling was proved to be more effective than general topic modeling in terms of packaging the social issues from the R&D perspective. Academically, research teams from various fields can take advantage of the extracted results to analyze current social issues and technology trends. Moreover, the social issues can not only be grouped from R&D perspective, but also can be grouped from other interesting perspective. This methodology can be further developed to discover different perspective of social issue packages. Practically, we also expect that our proposed methodology will contribute to establishing efficient R&D investment policies at the national level by enhancing the reusability of R&D knowledge. It should be noted that the same R&D technology can be related to multiple social issues. Thus, a news portal could also utilize the results by establishing an R&D perspective news package, which would allow news articles
to be provided according to different R&D indexes. Patent search services companies also can take advantage of our methodology to provide the patent information from the R&D perspective. It is possible to significantly reduce the burden of searching the patent by using different kind of keywords.

This study also had a few limitations. First, as the topic mapping results may vary depending on the target R&D domains, a future study is suggested to perform intensive analysis using different R&D domains. Second, the number of R&D terms in the selected criteria documents may vary according to their types and appropriateness. Therefore, guidelines for selecting criteria documents should be appropriately provided. Finally, because the R&D lexicon plays an important role in our approach, more rigorous consideration should be given to its refinement in order to improve the quality of results.

References

Aggarwal, C.C. and C. Zhai, “A Survey of Text Clustering Algorithms”, Mining Text Data, Springer, US, 2012, 77–128.

Agrawal, R. and M. Batra, “A Detailed Study on Text Mining Techniques”, *International Journal of Soft Computing and Engineering*, Vol.2, 2013, 2231–2307.

Aizawa, A., “An Information-Theoretic Perspective of TF-IDF Measures”, *Information Processing and Management*, Vol.39, No.1, 2003, 45–65.

Albright, R., “Taming Text with the SVD”, SAS Institute Inc., 2004.

Bae, J., J. Shon, and M. Song, “Analysis of Twitter for 2012 South Korea Presidential Election by Text Mining Techniques”, *Journal of Information Technology Applications and Management*, Vol.19, Vol.3, 2013, 141–156.

Barbakh, W.A., Y. Wu, and C. Fyfe, “Non-Standard Parameter Adaption for Exploratory Data Analysis” Springer, Berlin Heidelberg, 2009, 1–6.

Blei, D.M., A.Y. Ng, and M.I. Jordan, “Latent Dirichlet Allocation”, *Journal of Machine Learning Research*, Vol.3, 2003, Chicago, 993–1022.

Cai, D., X. He, and J. Han, “Locally Consistent Concept Factorization for Document Clustering”, *IEEE Transactions on Knowledge and Data Engineering*, Vol.23, No.6, 2011, 902–913.

Carnerud, D., “Exploration of Text Mining Methodology through Investigation of QMOD-ICQSS Proceedings”, in *QMOD-ICQSS Prague, Czech Republic*, Sep, 3–5, 2014.

Fan, W., W. Wallace, S. Rich, and Z. Zhang, “Tapping the Power of Text Mining”, *Communications of the ACM*, Vol.49, No.9, 2006, 76–82.

Feldman, R. and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, 2007.

Griffiths, T.L. and M. Steyvers, “Finding Scientific Topics”, *Proceedings of the National Academy of Sciences*, 2004, 5228–5235.

Han, J., M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd Edition, Morgan Kaufmann Publishers, 2011.

Hearst, M.A., “Untangling Text Data Mining”,
Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, 1999, 3–10.

Hofmann, T., “Unsupervised Learning by Probabilistic Latent Semantic Analysis”, Machine Learning, Vol.42, No.1–2, 2001, 177–196.

Holton, C., “Identifying Disgruntled Employee Systems Fraud Risk through Text Mining: A Simple Solution for a Multi–Billion Dollar Problem”, Decision Support Systems, Vol.46, No.4, 2009, 853–864.

Hong, D.S. and H.G. Kim, “From Big Data to Business Value: Emerging Technology Extraction and Analysis System Based on Korean Newspapers”, Journal of the Korea Entertainment Industry Association, Vol.8, No.4, 2014, 285–292.

Hyun, Y.J, H.J. Han, H.S. Choi, J.Y. Park, K.H. Lee, K.Y. Kwahk, and N.G. Kim, “Methodology using Text Analysis for Packaging R&D Information Services on Pending National Issues”, Journal of Information Technology Applications and Management, Vol.20, 2013, 231–257.

Kim, I., “The Value of Big Data and Strategy”, 2012 Big Data Search Analysis Technology Insight, 2012.

Kim, J.E., N.G. Kim, and Y.H. Cho, “User-Perspective Issue Clustering Using Multi-layered Two-Mode Network Analysis”, Journal of Intelligence and Information Systems, Vol.20, No.2, 2014, 93–107.

Nagaraj, R., V. Thiagarasu, and P. Vijayakumar, “A Novel Semantic Level Text Classification by Combining NLP and Thesaurus...
Concepts”, *IOSR Journal of Computer Engineering*, Vol.16, No.4, 2014, 14–26.
Newman, D.J. and S. Block, “Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper”, *Journal of the American Society for Information Science and Technology*, Vol.57, No.6, 2006, 753–767.
Ng, A.Y., M.I. Jordan, and Y. Weiss, “On Spectral Clustering : Analysis and an Algorithm”, *Advances in Neural Information Processing Systems*, Vol.2, 2002, 849–856.
Provost, F. and T. Fawcett, *Data Science for Business*, O’Reilly Media, 2013.
Romero, C., S. Ventura, and E. Garcia, “Data Mining in Course Management Systems: Moodle Case Study and Tutorial”, *Computer and Education*, Vol.51, No.1, 2008, 368–384.
Sebastiani, F., “Classification of Text, Automatic”, The Encyclopedia of Language and Linguistics, (ed.2), Vol.14, Elsevier Science Pub, 2006.
Stanvrianou, A., P. Andritsos, and N. Nicoloyannis, “Overview and Semantic Issues of Text Mining”, *ACM SIGMOD Record*, Vol.36, No.3, 2007, 23–24.
Sun, Y., “A Text Mining Approach to Analyze Public Media Science Coverage and Public Interest in Science”, *International Journal of Machine Learning and Computing*, Vol.4, No.6, 2014, 496.
Tseng, Y.H., C.J. Lin, and Y.I. Lin, “Text Mining Techniques for Patent Analysis”, *Information Processing and Management*, Vol.43, No.5, 2007, 1216–1247.
Wang, M.Y., D.S. Chang, and C.H. Kao, “Identifying Technology Trends for R&D Planning using TRIZ and Text Mining”, *R&D Management*, Vol.40, No.5, 2010, 491–509.
Wang, H. and Y. Ohsawa, “Innovation Support System for Creative Product Design Based on Chance Discovery”, *Expert Systems with Applications*, Vol.39, No.5, 2012, 4890–4897.
Weiss, S.M., N. Indurkhya, and T. Zhang, Fundamentals of Predictive Text Mining, Springer, London, 2012.
Wong, W.X.S. and N. Kim, “Reorganizing Social Issues from R&D Perspective Using Social Network Analysis”, *Journal of Information Technology Application and Management*, Vol.22, No.3, 2015, 83–103.
Xie, P. and E.P. Xing, “Integrating Document Clustering and Topic Modeling”, arXiv preprint arXiv : 1309.6874, 2013.
Xu, W. and Y. Gong, “Document Clustering by Concept Factorization”, *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004, 202–209.
Xu, W., X. Liu, and Y. Gong, “Document Clustering Based on Non-Negative Matrix Factorization”, *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, 267–273.
Yoon, J., “Detecting Weak Signals for Long-Term Business Opportunities Using Text Mining of Web News”, *Expert Systems with Applications*, Vol.39, No.16, 2012, 12543–12550.
About the Authors

William Xiu Shun Wong (williamwong@kookmin.ac.kr)

William Xiu Shun Wong received the B.S. degree in Computer Science from Universiti Sains Malaysia in 2011 and Master degree in Business IT from Kookmin University in 2015. He is a Ph.D. candidate in Business IT at Kookmin University. His current research interests include text mining, data mining, and text classification.

Namgyu Kim (ngkim@kookmin.ac.kr)

Professor Namgyu Kim received the B.S. degree in Computer Engineering from Seoul National University in 1998 and Ph.D. degree in Management Engineering from Korea Advanced Institute of Science and Technology (KAIST) in 2007. He has been working for Kookmin University since then. His current research interests include text mining, data mining, and data modeling.