Information Processing and Management 56 (2019) 1506–1525

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Can media forecast technological progress?: A text-mining approach to the on-line newspaper and blog's representation of prospective industrial technologies

Leo Kim[*,a], Jaewook Ju[b]

[a] Ars Praxia, The K-Twin Tower 2F, Jongro-gu Jongro 1 Road 50, Seoul 03142, Republic of Korea
[b] Seoul Institute, Nambusunhwan-ro 340-gil, Seocho-gu, Seoul, Korea

ARTICLE INFO

Keywords:
Industrial technology
Text mining
Semantic network analysis
Technology prediction
Social media analysis

ABSTRACT

This article critically assesses the utility of both personal blog and mass media (on-line newspaper)'s coverage of future technology in forecasting the prospect of industrial technology. By analyzing the statistical pattern of the South Korean media salience in 13 novel industrial technologies in 2015, the study argues that the mass media is more biased than aggregated blogs in depicting promising new technology. In doing so, the authors present a methodical pathway to address the bias with acceleration and skewness index. The article also applies semantic network analysis for the collected textual data of blogs to represent and interpret people's perspective of industrial technologies. By extracting key concepts from PBS index and merging semantic networks of the 13 technologies, the study derives a key insight that nurturing expertise in the coming era of artificial intelligence and robot would become crucial to integrate key technological competence.

1. Introduction[1]

The digital economy has transformed the economic environment significantly. The term "digital economy" means an economy based on information and communication technology (hereafter ICT) where at least one of the traditional economic elements (i.e., productive agents, commodities, service, market, currency) has been replaced with digital computing technology, creating new economic value. Whereas the process of production in the past was only partly aided by computer-operated machine or information technology, the current economy represents a structure where the ICT has evolved to an end goods as well as intermediary goods just as software and digital contents are rigorously transacted as valuable commodities or services. Nowadays, the utilization of the ICT is not confined to a narrow domain of computer, software or electronic goods, but practically permeate in every area of industry including production, sales, operation and service to profoundly transform the characteristic of economic activity itself (OECD, 2017: 11–13).

In order to predict and forecast economic trajectory, understanding the evolving pattern of digital economy is gaining much importance. While the boundary between the traditional and digital economy is getting blurred, the growing proportion and depth of the latter in contributing to economic mechanism is evident. However, capturing the actual digital activity is less clear than expected as the predefined categories of economic activity is still based on the traditional perspective, and statistical application based on those

* Corresponding author.
E-mail address: leo.kim.praxis@gmail.com (L. Kim).
[1] This research was funded by the Korea Information Society Develoment Institute (KISDI) and completed in February 2016

https://doi.org/10.1016/j.ipm.2018.10.017
Received 10 July 2018; Received in revised form 30 September 2018; Accepted 22 October 2018
Available online 05 January 2019
0306-4573/

categories is not relevant to represent the extremely rapid and microscopic nature of activities in digital economy. Therefore, we need an alternative methodological approach.

Internet data has recently been regarded as a viable source to effectively represent the diverse human activities, including economic, in magnanimous volume. Social media, on-line newspapers, and personal blogs are producing infinitesimal messages that are described in text, image, and video data. If the researcher knows how to derive meaningful information out of the data effectively, we will be equipped with more powerful apparatus for social forecast. In reality, the internet data would be “overwhelmed” by unnecessary noise as well as various sets of bias, which would exaggerate the studied outcome and mislead the researcher. Hence, we need to find a way of critical (re)analysis of those media data.

In this article, the authors have attempted to utilize a text-mining methodology to predict and interpret the industrial technological movement. Initiated by the Korea Information Development Institute (KISDI), the research scheme and target technologies were selected by the organized expert panel (led by Dr. Jaewook Ju) while also considering their relevance to the governmental interest to promote key area of emerging industry. The selected thirteen area of emerging industrial technologies include: Internet of Things (IoT), electric vehicle, autonomous vehicle, fintech, drone, artificial intelligence (AI), robot, next generation battery, healthcare, wearable, 3D printer, virtual reality (VR), and big data. The authors followed the selection of these thirteen areas after several times of the panel discussion at KISDI, under the “Methods for Data-Based Forecasting Digital Economy” scheme by which this research was funded. The panel members from various ICT domains tried to include presumably promising new areas such as next generation battery or drones in addition to the areas already widely mentioned to by expand the scope of the analysis.

In doing so, the authors critically reassessed how the frequency of news coverage alone might mislead technological expectations. As we will discuss, some of the areas we studied reveal a concentrated pattern of news coverage that is mainly dependent on the domestic companies’ release or marketing of new industrial products. Alternatively, we propose to utilize statistical indices such as acceleration and skewness to overcome the bias. In order to better reflect the technological trend, the combined utilization of statistical analysis of news and the operation of “Windrose Model” (Bauer & Gaskell, 2008) by semantic network interpretation of blog data turns out to be crucial.

2. Methodology

2.1. Background of research

The mainstream of future studies faces a drastic change around the period of the Great Depression and the Second World War. If the epistemological background in the previous period represented a belief in rational thoughts, future studies afterwards recognize the limited nature of objective thinking and growingly emphasize the subtle dynamics in which “irrational” elements like personal experience and perception, cultural relationships play a wider role. The current future studies have been initiated to predict future technological trends, gradually expanding its domain from state affairs to corporate operations, inter-governmental relations, NGO activities, etc.

Diverse forecasting methods have been developed since Kahn and those were systemized by Glenn and Gordon (2003) and Georghiou (2008). Georghiou categorized the existing methods in terms of three conceptual axes, (1) time, (2) characteristics of results, and (3) use by area. According to Georghiou, the methods are divided into exploratory and normative ones based on time and into qualitative, quantitative and hybrid ones based on the characteristics of the results. Exploratory forecasting methods are the ways to forecast the future based on the knowledge that we currently know. They include Delphi technique, cross-impact analysis and trend extrapolation. Normative forecasting methods are those that design the realization plan based on supposedly desirable image of the future, including scenario workshop, relevance trees, morphological analysis and technology road-mapping. Qualitative methods utilize the intuition, insight, expertise and decision of expert, which include scenario workshop, SWOT, futures wheel and qualitative trend analysis. Quantitative methods use statistical numbers and models that include trend extrapolation, modelling and bibliometric trends. Quantitative methods are advantageous in that they exclude subjectivity of researcher as much as possible and have logical objectivity, but they have a disadvantage that it is difficult to perform flexible and in-depth analysis because they are based on very limited form of data. Hybrid methods apply mathematical and statistical principle to quantize subjective knowledge and viewpoint of experts, which include Delphi technique, cross-impact analysis, structural analysis and technical road-mapping.

The text-mining method we propose is an eclectic approach that is positioned between quantitative and qualitative methods. While the prediction of technological growth based on statistical representation of extracted keyword distribution is quantitative, the interpretation of possible key factors for such growth relies on people's discourses of technology. The underlying assumption is that any scientific knowledge and technological progress is dependent on the “co-production” (Jasanoff, 2006) between the expert's technical improvement and the society's recognition of its significance. While there are a few cases that adopted on-line text-mining approach (Li & Wu, 2010; Rickman & Cosenza, 2013; Vecchio, Mele, Ndou, & Secundo, 2018), the “word of mouth” of on-line media in their research were interpreted rather uncritically to forecast the trend of designated industry. In comparison, we try to propose ways to critically assess the characteristics of on-line buzz, and filter-out core elements for restructuration of data and rigorous interpretation with theoretical (mainly from science technology studies and social psychology) considerations.

The extracted core frame of on-line news and people's blogs respectively exhibit the experts’ and the lay actors’ activity, evaluation, and expectation of novel technology. The relative advantage of text-mining is that it can handle large sizes of data promptly to forecast the future trajectory of salient technological field; and also evaluate their current characteristics from people's representations. Our intermediary approach to integrate quantitative/qualitative research and expert/lay perceptions aims to induce reliable expectation and interpretation. And this requires a standardized process to manipulate the big data. Then, the question arises:

How to filter out, restructure, and extract the core significations reliably with the big data analytic method? Following sections try to answer this question.

2.2. Statistical indices for critical forecast

Term frequency is a classical way to retrieve information from a large amount of text (Lent, Agrawal, & Srikant, 1997, 1997; Feldman et al., 1998). It is still used for recent studies of predicting market (Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2014; Sun, Lachanski, & Fabozzi, 2016) or forecasting hospital admission of emergency department (Lucini et al., 2017) as a basic statistic. Term frequency gives much information of raw texts. However, a systematic comparison between specified keywords from newspapers requires more than the frequency of keywords. In other words, there is little criteria how much difference between term frequencies implies that one keyword is more promising than others. For example, we cannot simply conclude that “3D printer” is more promising technology than “Internet of Things” from the fact that frequency of “3D printer” occurred at the newspaper is twice as much as “Internet of Things.”

To overcome such difficulty, we propose to use two major concepts: acceleration and skewness. The term acceleration is translated from the notion of physics, and the skewness is from statistics. Note that difference of term frequencies has given information that the term drew more attention (or less attention) in base period. However, difference of term frequencies of a word is not directly comparable with that of the other word. For example, suppose that “3D printer” has term frequencies 1000 and 1100 for each time T_1 and T_2, and “Internet of Thing” has term frequencies 100 and 150 for each time. We cannot conclude that “3D printer” drew more attention than “Internet of Things” because its difference is 100, which is greater than difference of frequencies of “Internet of Things,” which is 50. This is because it changed only 10% as a relative portion, which is far less than the 50% of “Internet of Things.” On the other hand, we cannot conclude that “Internet of Things” drew more attention than “3D printer” because of the change of relative portions; still term frequency of “Internet of Things” is much smaller than that of “3D printer.”

Instead of comparing differences directly, we can evaluate words by their potential prospect. To define this value, we adopted basic model of physics: potential, distance and velocity. We assume that the occurrence of term can be viewed as traveling of term on a space consisting of media. Then, this occurrence can be viewed as a distance the term travels on the space at specific time interval. Distance traveled by term is proxy of having attention from media. From this allegorical perspective, we can translate the meaning of acceleration of term frequencies as a potential prospect of attention the term will have.

Thus, acceleration of term frequencies is defined as the second derivative of term frequency, a function of time. This acceleration can be defined for each time period or defined for whole time period. To compare overall tendency of keywords, we can assume that potential of term frequency is constant during whole time. By this, we obtain the acceleration by second order polynomial regression of term frequencies. Note that acceleration means a power of the movement of the keyword by Newton's second law. The higher acceleration, the more increasing speed of keywords frequency, which implies that more (potential) attention is expected in the future during the observed period.

Skewness is a measure of the asymmetry of the distribution (Hogg & Craig, 1995). We assume the distribution of term frequency comes from a probability distribution with probability variable X. Then, given term frequency data can be regarded sample space of X. Then skewness is defined as the expectation of the cube of X minus average divided by standard deviation.[2] From this measure, we can get the stability of attention on a keyword. For example, if a distribution of keyword frequency has been more skewed than the other, then this implies that lots of frequencies occur at a limited time period. Negatively skewed implies that period which gives high term frequencies occur at a small time period compared to whole time period. Conversely, positively skewed implies that period which gives low term frequencies occur at a small time period compared to whole time period. In any case, term frequencies are not stable if we define consistency of term frequency as how much the given distribution differs from the normal distribution. What the authors try to verify is if a term is “bursty” (not stable) in short period. In this regard, we use absolute value of skewness as index of stability (Lee, Lee, Kim, & Park, 2012).

In summary, keyword frequency is used as a basic statistic in text mining. However, we can draw further information by tree indices, acceleration, relative standard deviation, and skewness from term frequency. Acceleration gives how much power of momentum a keyword increases, from a near future perspective, in media representation. Relative standard deviation and skewness measure how much a trend of keyword frequency is consistent. If we assume that media reflects some prospect of technology, these indices can be used for forecasting whether and which technology is promising or not.

2.3. Extracting potentially influential concepts in semantic network

In order to represent the discourses of news and blogs in a systematic form, we utilized semantic network analysis to interpret the linkage pattern of keywords that was considered to reflect the typical frame of lay perception, and to extract the central organizing idea of the semantic network (Gamson & Modigliani, 1987; Kim, 2013). Semantic network analysis is a form of content analysis which extracts the network of relations between objects as expressed in a text, in order to represent a discursive model as a visible map (Carley, 1993).

The selective links of concepts in the semantic network represent a symptom of the social representation (Moscovici, 2000) of

[2] Skewness is defined as $E\,[(\frac{X-\mu}{\sigma})^3]$, where $E[Y]$ denote the expectation of given random variable, μ denotes $E[X]$, and σ denotes standard deviation of the ramdom variable X.

utterers, and they are incorporated into discourse analyses that delve into microscopic relations of power among actors – mediated by language.[3] Social representation theory (Deaux & Philogène, 2001; Moscovici, 2000) presents a formal way of considering multiple levels of signification in communication, by actively incorporating both experts' and lay people's knowledge and perception.

By focusing on the pragmatic and contextual nature of sense making, a structural aspect of signification can be methodologically captured from the distribution of words internal to a large text corpus produced by the members of a culture (Suerdem, 2013). According to Bauer and Gaskell (2008: 345), the social representation is depicted as a "centripetal intentionality of different communities towards the common referent" whose petals of the "windrose" (Fig 1) are "social milieus, schematised as triangles of mediation, the particular milieus of communication making reference to" a certain technology.

The analyzed discourse by the semantic network analysis with the translation of the Windrose Model can effectively explain how the socially controversial issues are being defined, what are the salient causal interpretations, what are the associated value references and what could be the converging expectation in the future signified by "central organizing idea" (Gamson & Modigliani, 1987). The research object in relation to technological forecasting can be described as a latent "cloud seed", functioning as a constraint of reality (see Fig. 1), of the windrose. Normally, to represent the "central organizing idea" in the network, various centrality indices could be used (Kwon, Barnett, & Chen, 2009). For our specific purpose of identifying "cloud seed" and future prediction, we need to focus on keywords that are not already mostly influential, but have a potential quality to become so. In other words, numerical representation of latency is required.

From the perspective of semantic network, interaction between two nonadjacent nodes of concepts is likely to depend on another concept for reference that functions as a catalysis to join metalanguages of concepts. This function of salient "denotation" (Barthes, 1967) is translated into a node with the highest betweenness centrality (Freeman, 1979) in the semantic network, when the keyword lies on the paths between the trigger of information and referent, performing a mediating role as a semiological facilitator of communication (Kim, 2013). The function of "boundary spanning" in the social network analysis is similar to the mediating role in which nodes with high betweenness centrality perform, but focus on more specific role to facilitate the flow of information between nodes that either have no physical or cognitive access to one another, mainly composed of isolated clusters in need of connectivity (Long, Cunningham, & Braithwaite, 2013). Numerically, this characteristics of potential boundary spanner (PBS) is represented by dividing the betweeness centrality by degree centrality in the network.

For the stable representation of such potential influence, the number of nodes in the network should be controlled in a standardized way because centrality indices, including betweenness and degree, vary according to the size of the network. For the standardized extraction of a core network that is composed of homogeneous number of nodes, the backbone extraction model (Serrano, Boguñá, & Vespignani, 2009) was adopted. An important aspect of this model is that the ensuing reduction algorithm does not belittle small nodes and links in terms of frequency while offering a stable automatic procedure to reduce the number of connections by taking into account all of the scales – the representative reduction of the original network (Kim & Kim, 2015).

2.4. Research process

To derive a consistent outcome without human intervention in coding a textual data, utilization of an automatic algorithm based on the aforementioned assumption becomes vital. Based on the methodological assumptions and procedure, the keywords in the text were coded and analyzed automatically by the commercial text-mining tool *Optimind* (ver. 2.0).[4] *Optimind* is an automatic semantic network tool, based on the natural language processing, automatic coding of textual data with hidden Markov model and backbone extraction of the semantic network (see Appendix I). By operating the computerized system, the statistical analysis and text mining go through the following stages in Table 1.

We collected data of Korean on-line news and blogs from the portal website Naver (www.naver.com) with designated search words mentioned in Table 1. 1) c. The Korean media portal Naver occupies 83.3% of search records, followed by another domestic search engine Daum (13.0%) and Google (4.3%) in 2015.[5] Currently, the consumption of on-line newspaper is also predominated by Naver, inducing 13 million daily users or 66.3% of on-line news readers entirely (Kyunghyang Biz, 25 April 2018). Given the 3 quarter period of data collection (from 1 January to 30 September 2015), the frequency of news documents were measured in daily, monthly, and quarterly basis. As mentioned in the methodology section, acceleration was used to predict the future salience of news coverage and skewness to account for the impartiality in comparison to the frequency that had been traditionally used to represent media salience (Bauer & Petkova, 2005).

To further understand the people's perception and formalized frame of technological movement, we utilized blogs written in Korean and applied semantic network analysis. In comparison to existing methodological works (Jurowetzki & Hain, 2014; Kim & Lee, 2015) that either focused on network density or extraction of community structure with a data-mining algorithm, our study tried to present a standardized way to combine basic statistical measurement and the utilization of PBS in extracted semantic network in

[3] It should be noted that the early proponents of social representation theory like Serge Moscovici paid attention to the underlying social groups of particular representation. The authors consider the analytical approach in the mid 20th century rather simple to grasp the multifaceted and highly overlapping social identity in contemporary era. Instead, reversing the original theory and applying up-to-date text mining method, we try to extract the "converging" aspect of representation despite heterogeneous social backgrounds of utterers.

[4] Designed by social network analysis company Ars Praxia (http://www.arspraxia.com). For the academic references utilizing the tool, see Kim and Jang (2018) and Kim and Kim (2015).

[5] Source: Internet trend (http://internettrend.co.kr)

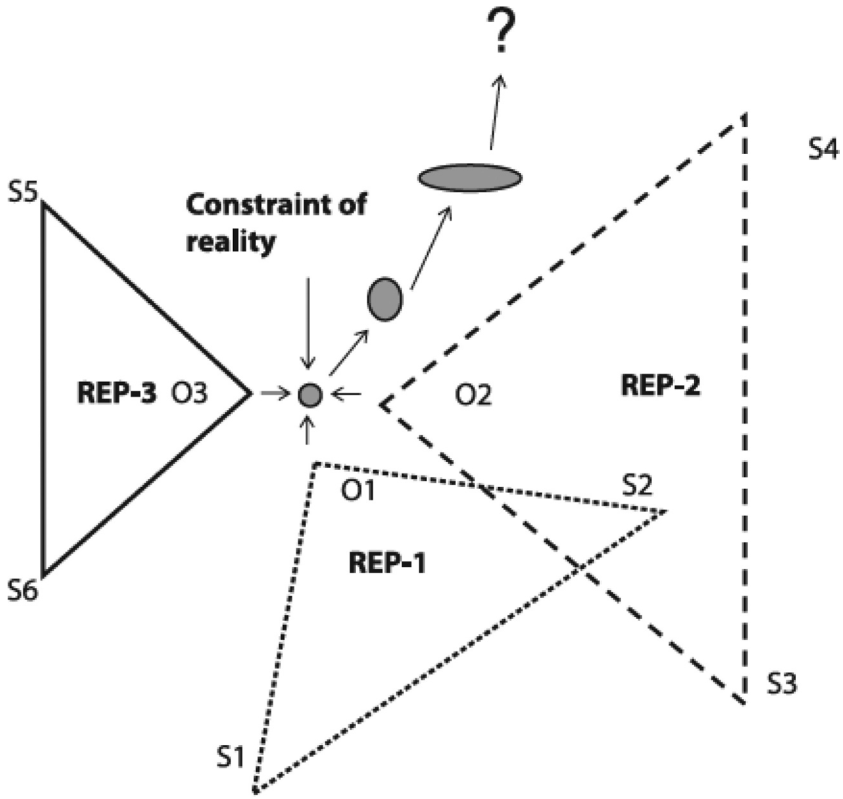


Fig. 1. The Windrose Model (Source: Bauer & Gaskell, 2008).

Table 1
Data-mining process.

1) Collection of data:
a. Target: Korean on-line news and blogs in the portal Naver
b. Period: 1 January 2015 ~ 30 September 2015
c. Search word: Firstly, we applied Korean terms for “electric vehicle”, “autonomous vehicle”, “fintech”, “drone”, “AI”, “robot”, “battery”, “healthcare”, “wearable”, “3D printer”, “VR”, and “big data”. Secondly, we filtered out irrelevant documents by applying various delete words.

2) Statistical analysis:
a. Quarterly graph of technological trends during 1 January ~ 30 September 2015 with the frequency of news documents in each field
b. Applying acceleration and skewness based on the daily frequency of documents
c. Binary comparison between average frequency and additional measures of accleration and skewness and interpretation of the plotted graphes

3) Semantic network analysis
a. PBS measurement of the extracted blog network in each technological field
b. Merging 13 blog networks by restructuring each technological field by firstly organizing a network with technological title connected with top 20 keywords that have highest PBS, and secondly amalgamate the 13 networks into one network
c. Tagging associated nodes with technology in the single network with 3 aspects: technological, industrial, and social

4) Interpretation
a. Systemic forecast of technological trends and qualitative review
b. Interpretation of key factors that would be significant in the progress of overall technological field based on the merged network feature and tagged keywords that have latent influence

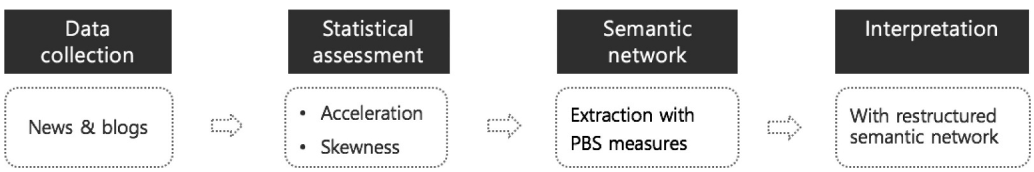


Fig. 2. Overall flow of research.

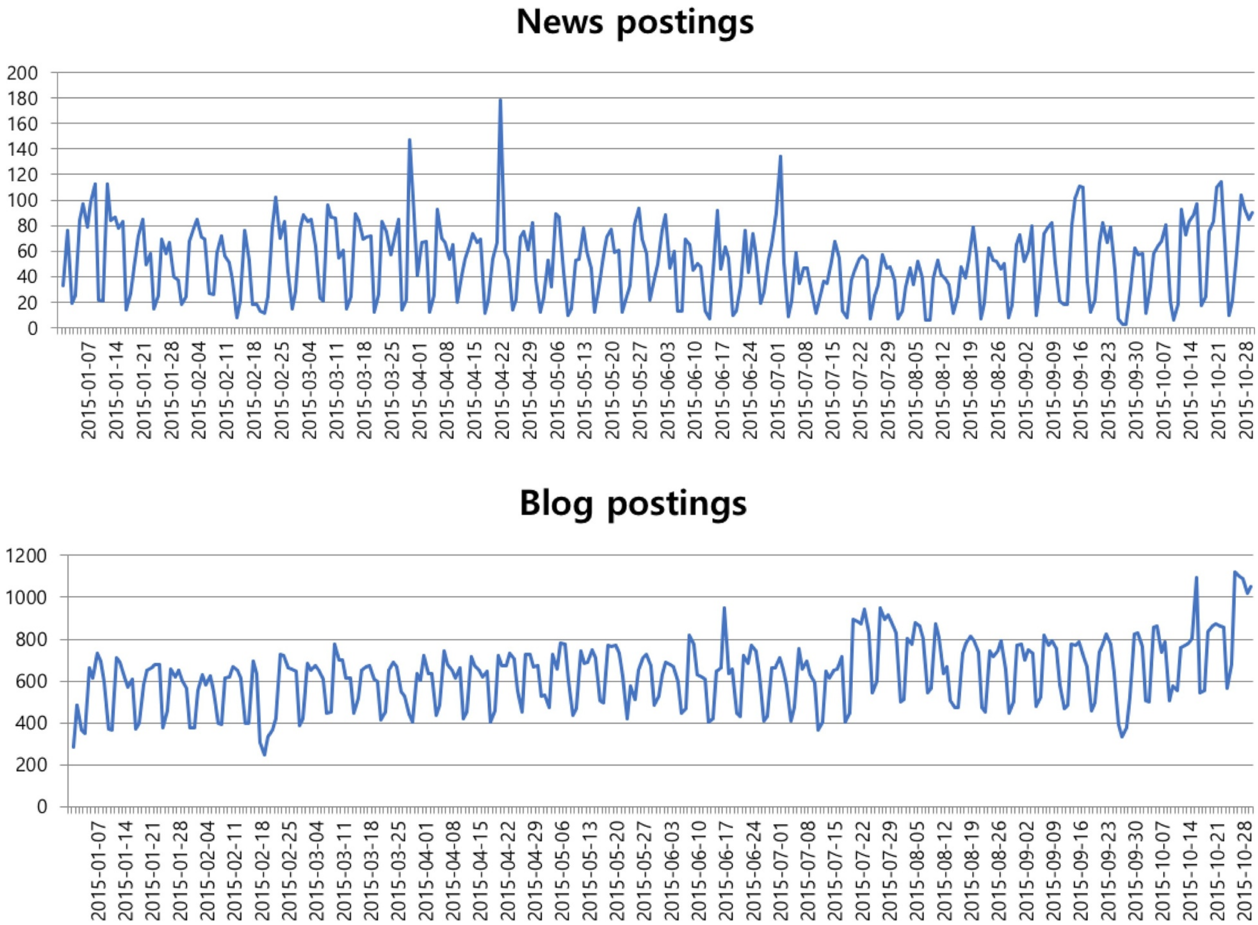


Fig. 3. Basic Statistics (overall).

technological forecasting. In addition, as shown in the result, our way of integrating disparate networks by reformulating core features of each network proposes a novel pathway to reformulate semantic network in order to detect and represent emergent characteristics of socio-technological interaction.

Overall, our research flow (Fig. 2) essentially aims to present an enhanced method for critical assessment of data, and actively interpreted the latent properties in the textual big data analysis. If the statistical assessment provides a guideline to filter prospective technologies, the following semantic network analysis concerns on a societal eco-system and mediating concepts that promote technological innovation. Extracted key concepts with data analysis are particularly designed to enrich the process of qualitative interpretation that is aided by formalized methods to manipulate large sized data in both efficient and comprehensive way.

3. Results

3.1. Critical review of statistical results

Overall, there were 15,615 postings of news in Naver from 1 January to 30 September in 2015. The average daily postings were 51.5. As shown in Fig. 3, the fluctuation of postings was remarkable, ranging from 10 to 180 daily. Blog postings were also collected from Naver platform especially for more detailed content analysis, which amounted to 191,901 in total or 633.3 postings daily. Apparently, the postings of Blog outnumber that of news reporting, and generally contain more genuine expressions and opinions of technologies with lengthy texts.

Fig. 4 shows the quarterly buzz of 13 technologies in news reporting and blog postings. In the news postings, healthcare (top) shows a drastic decline in the third quarter while IoT and robot (second and third top) bounces in the same period. For the Blog, robot is predominantly higher than other subjects. Most of the subject dealt with comments on recent movies like Ex Machina that popularized fear of androids that is undifferentiated from and smarter than humans. The second top was drone that recently drew much public attention for new usability.

If news coverage adequately indicates the intensity of controversy and official interest in scientific issues (Kim, 2011), personal blogs represent lay people's interest and imagination of issues that might be partly framed by institutional media. Here, the framing

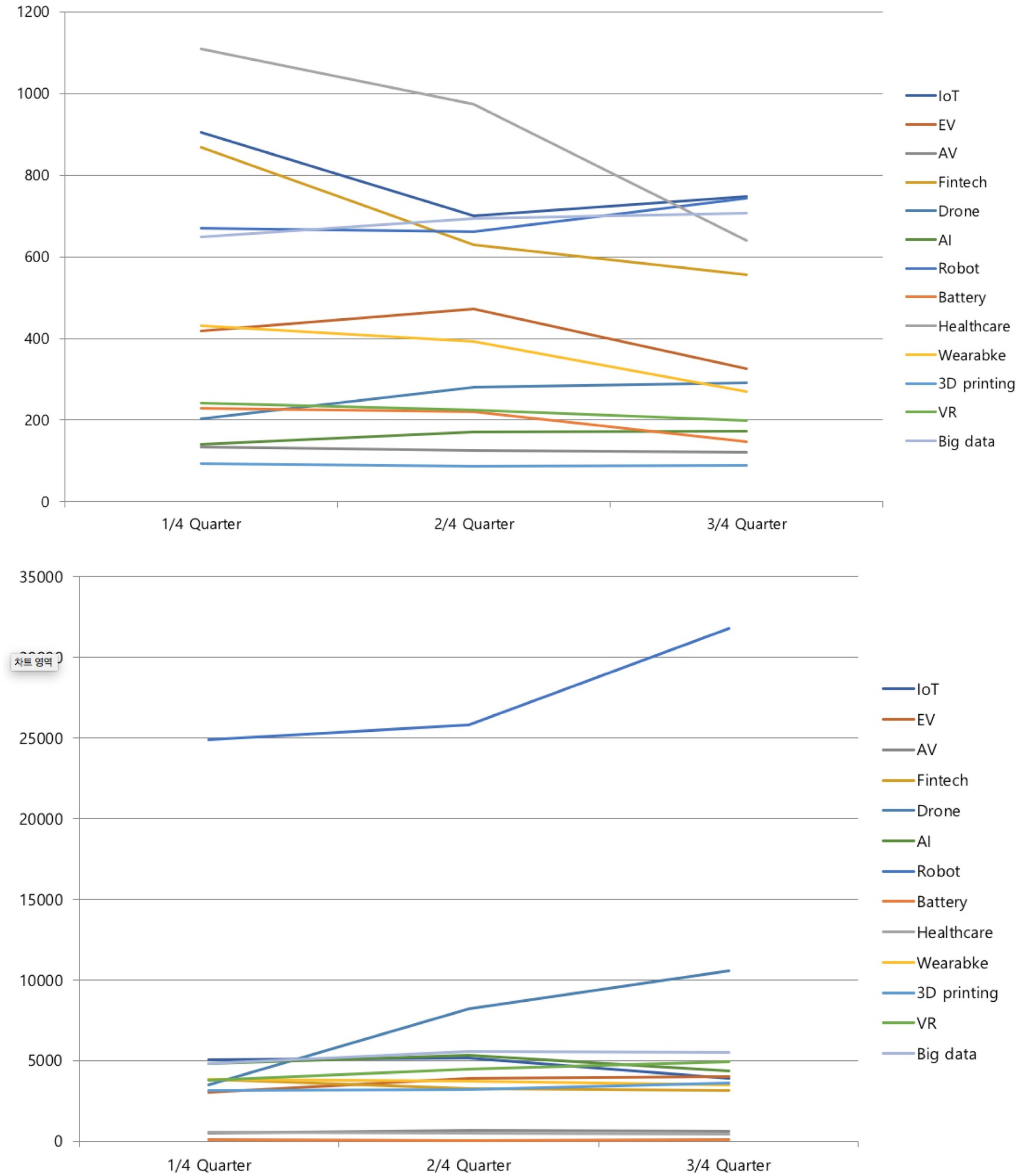
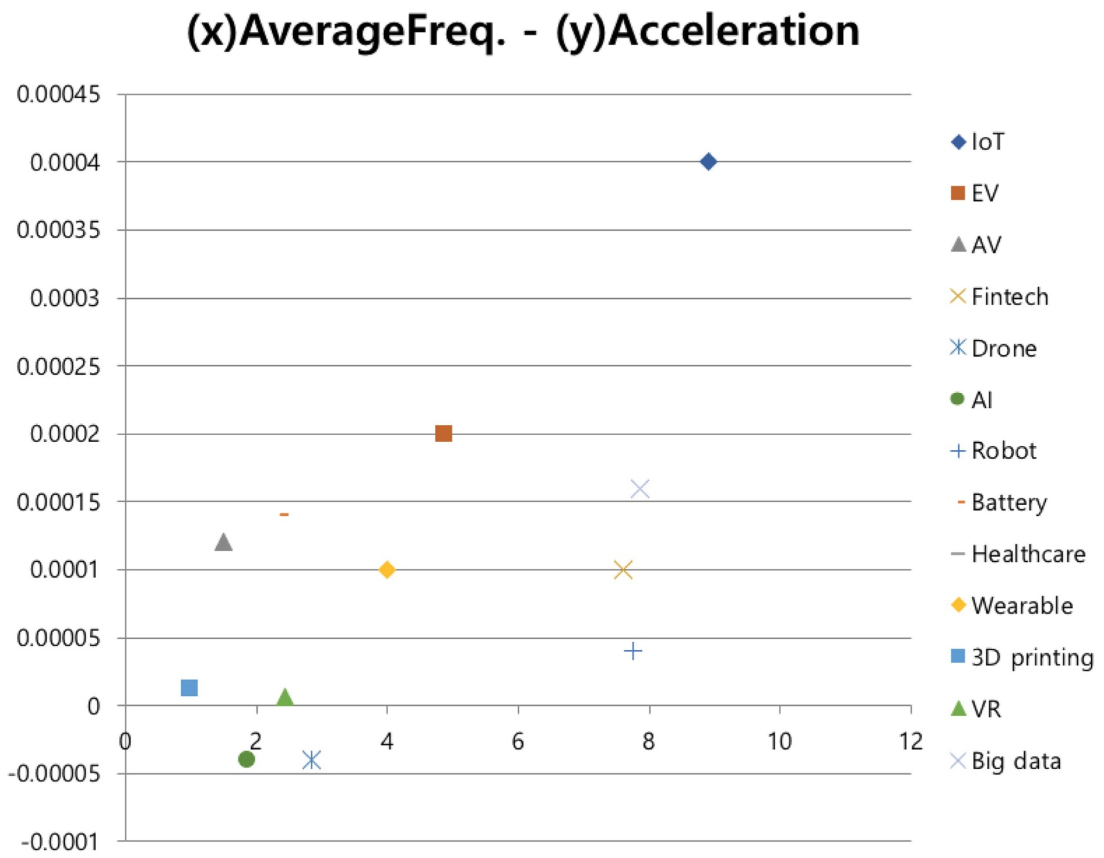


Fig. 4. Quarterly buzz of news comments and blog postings.

refers to "the way a story is told by unfolding arguments, using metaphors and imagery that define a problem, arriving at causal or moral attributions, and prescribing particular remedies" (Entman, 1993: 52). While both of the media platforms have their own frames, official newspaper usually precedes in shaping public issues and argument. Therefore, the structural pattern of news media frame requires a critical review with more sophisticated statistical analysis.

From the critical perspective, Fig. 5 depicts the distribution of each technology according to its average frequency, acceleration and skewness based on daily postings (see Appendix II and III). In the Korean newspaper, IoT appears to have the highest position in



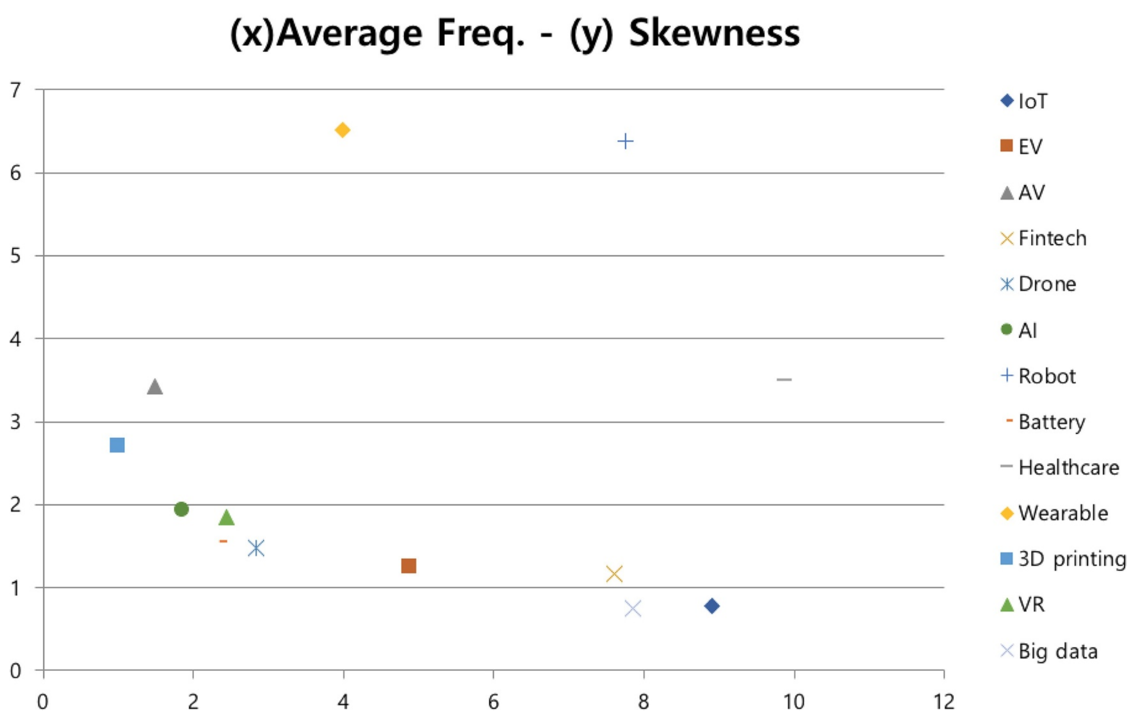


Fig. 5. Statistical analysis of news postings.

terms of both average frequency and acceleration. This means that IoT is currently receiving the highest attention and is likely to continue to draw much interest in the near future. Big data occupied the second position in average frequency, and the 3rd in acceleration. Although positioned in the 5th and the 7th respectively, electric vehicle ("EV") and new-generation battery ("battery") ranked the second and the 4th in acceleration, just beneath the much popularized IoT and big data. Considering that electric vehicle was met with some skepticism and the biggest South Korean car manufacturer, Hyundai Motors, preferred to developing fuel cell (hydrogen) vehicles in 2015, this latency seems to reveal an interesting aspect. On the other hand, AI and drone marked the lowest position in terms of acceleration. The feature was considered to reflect the lack of domestic industry where core technologies are overshadowed by the US and Japan (AI) or China (drone).

In addition, the skewness graph reveals that wearable and robot are situated in the top of *Y* axis, meaning that their media representations are highly concentrated in a short time period and more volatile than other technologies. After reading those news contents manually, it turned out that most of the news coverage either reported the launch of new wearable products of Korean electronics companies or consisted of reviews of robot related movies. In short, AI and robot were not considered a socially valuable news agenda from those official media whereas big data, IoT, and fintech (from the bottom) was viewed with relatively constant interest.

3.2. Semantic network analysis of blogs

As witnessed in the previous analysis, media representation of technology is not without bias. In some ways, it is an "institutionalized bias" because marketing strategy of big companies could prevail over any genuine discourses. In this regard, the social representation of lay peoples' blogs could be a more interesting representation of social thoughts. If newspaper requires regularity and consistence, the volatile characteristic of personal blog reflects the dynamic mindsets that contextualize (Bauer & Gaskell, 2008; Jasanoff, 2006) lay people's willingness to accept a certain technology.

In order to extract important keywords of bloggers, potential boundary spanner (PBS) measurement was adopted to extract some top-listed concepts in each technology. More familiar centrality indices in social network analysis such as degree, closeness and betweenness (Freeman, 1979) highlight a certain "status quo" nodes that are already salient in the network. The limitation of these measures is that it is more or less dependent on the frequency of connections while "latency" that is significant in future forecast reflect critical location rather than already salient frequency. In this regard, PBS might reflect an efficient location of node that could potentially function as a critical passage point.

Table 2 is the result of the extracted top 20 keywords along the PBS score in each technological domain. Based these top-listed keywords in terms of latency, we proceed to reformulate and integrate 13 separate networks to capture an idea of entire eco-system. To canvass some salient characteristics, 3D printing was mainly represented with medical technologies ("artificial hand") requiring new feature of "design" by "convergence" and new enterprise ("founding"). Drone was related with visual functions ("camera", "visual") and military use ("military). Robot was strongly related with technologies dealing with artificial intelligence ("intelligence", "system", "computer", etc.) with anticipation of future development. (Next generation) Battery was predominantly related with electric vehicle ("EV") and both the role of government and pioneering firm like Tesla was emphasized. In regard to big data, "accumulation" and "measurement" acquired the highest score in PBS. Emphasis of "business" and "operation" followed. Wearable consisted of products that are already in the market, with some anticipation to link with "healthcare" or new form of "service." For the AI, the countries of Japan, US, and UK were explicitly included in the top score sheet. The issue that was closely connected with autonomous vehicle (AV) was "safety." AV included many terms that were also salient in robot, AI and big data such as "control", "operation" and "information." Electric vehicle was linked with environmental issues ("eco-friendly") coupled with the scandal ("fiasco") of Volkswagen's illegitimate report of fuel efficiency. Fintech contains a number of keywords of fintech services, but the firstly ranked work was "information." Novel healthcare technology and Internet of Things (IoT), likewise, demonstrate much emphasis on "external data", "information" and "big data". Finally, virtual reality (VR) reflects people's expectation of immersive experience ("immersion"), new type of gaming activities ("game") and movie-watching ("movie").

3.3. Integration of technological networks and hidden agenda

As witnessed in the previous section, a number of top listed keywords overlap across the 13 technologies. When conducting semantic network analysis, researchers usually bring out a complex picture of overall network by processing each textual corpus or merged one entirely. In order to explore what kind of mutual relations could emerge from the 13 separate corpuses, integrating the texts and analyzing them together is vital. The conventional way of conducting one-mode network analysis by simply merging and analyzing 13 corpuses casts a few limitations: (a) The integrated network is overly complex and unrecognizable. (b) When each technology is initially supposed to be given equal attention, simple amalgamation of text over-represents keywords that have high frequency even when the frequency is not an adequate index to oversee relational properties of the network. (c) The single dimensional (one-mode) network of processed keywords in most cases nullifies the relation between the indexed technologies and connected keywords.

To overcome the limitations, we utilized a two-mode network analysis (Bargatti & Everett, 1997) with a modified format. The researchers firstly drew a line between the 13 technology labels (first mode) and following top 20 keywords (second mode) with PBS in each technology. Secondly, we created a direct line between the technology label and enlisted keywords, forming a new egocentric network. Thirdly, the 13 sets of reassembled egocentric network was merged to represent an ecosystem of various technologies that are connected by overlapping keywords. Finally, we recalculated the PBS measure in the integrated network to find out potentially

Table 2
Top 20 keywords with PBS.

Rank	3D printing	Drone	Robot	Battery	Big data	Wearable	AI	AV	EV	Fintech	Healthcare	IoT	VR
1	kinds	drone	awesome	EV	accumulation	base	Japan	realization	eco-friendly	infotmation	result	information	immersion
2	object	advantage	control	share	measure	Android	control	safety	hybrid vehicle	insurance	external data	collection	hologram
3	method	special	completion	charge	business	vibration	level	semoconductor	fiasco	feature	model	production	game
4	artificial hand	experience	cat	firm	operation	check	growth	possible	time	crowd funding	trend	telecommunication	matrix
5	sector	driving	Japan	possible	feature	technology	US	automobile	Germany	securities	characteristics	product	realization
6	business	camera	technology	quality	precision	realization	UK	road	Samsung SDI	card	collection	connection	movie
7	design	purchase	intelligence	brand	kinds	sell	design	robot	internal	base	reference	device	VR
8	various	future	US	government	collection	real	scout	utilization	sales	investment	items	startup	device
9	domain	military	development	utliization	video	innovation	Dion	control	Nissan	action	medical	object	reality
10	file	Inspire	feature	core part	realtime	function	real	check	necessary	software	production	big data	character
11	metal	utilization	system	chemical	large size	healthcare	appearance	operation	patent	transfer	firm	technology	research
12	printing	addition	making	business	Seoul	product	movie	run	Volkwagen	characteristics	product	data	utilization
13	market	function	face	automobile	creative economy	market	responsibiily	regulation	US	items	envision	innovation	technology
14	modeling	feature	kid	Samsung SDI	economy	expert	sensibility	test	business	mobile payment	investment	economy	market
15	convergence	video	research	industry	solution	purchase	programing	improvement	system	finance	reference	security	developer
16	industry	possible	computer	battery	IoT	iphone	developer	utilization	operation	paycheck	guarantee	solution	platform
17	founding	pictures	human	mode	information	economy	love	information	market	firm	customer	applications	innovation
18	hardware	booth	machine	market	data	utilization	emotion	speed	Japan	perception	draft	hardware	device
19	growth	safety	future	occupation	capability	service	expert	charge	subsidy	preference	transfer	creative economy	economy
20	certificate	price	utilization	Tesla	necessary	creation	genius	assist	industry	market	research	system	fantasy

Table 3
Top 30 keywords arranged by T-I-S category.

Technology	PBS	Industry	PBS	Society	PBS
System	0.349392	**Service**	0.358367	**Expert**	0.368141
AI	0.212077	Parts	0.193248	Innovation	0.236248
Software	0.205635	Automobile	0.077601	Certificate	0.078214
AV	0.205349	Brand	0.037259		
Video	0.201188				
Fintech	0.194322				
Platform	0.188102				
Drone	0.187052				
3D printing	0.175195				
VR	0.174688				
Battery	0.168318				
Big data	0.159367				
Robot	0.156885				
EV	0.136954				
IoT	0.132908				
Healthcare	0.109950				
Material	0.106667				
Wearable	0.106507				
Solution	0.101953				
Hardware	0.062429				
Charge	0.045694				
Control	0.045108				
Data	0.037259				

prominent “obligatory passage points” (Callon, Law, & Rip, 1986; Latour, 2005) in the Actor-Network.

From the integrated network of keywords (originally in Korean: see Appendix IV), 30 keywords were extracted that were enlisted in the Top 30 of potential boundary spanner measure. The keywords were tagged by three characteristics, namely, [technology], [industry] and [social] in order to study how these properties are manifested in bloggers’ frames of technology and potentially influence on the interaction in the technological field. In doing so, the authors attempted to extract key mediating concepts not only from technological domain per se, but also related arena of business and society.

Table 3 shows the distribution of those keywords that are arranged by PBS order in each column. All of the 13 technology labels were included in [technology] category, and “system” was ranked at the top of the list. The reason the research panel found was that “system” was connected to a number of core technologies as observed in Table 4a. In [industry] category, 4 keywords emerged. From the perspective of latency, “service” turns out to be the most salient concern. This indicates that a number of emerging technologies are yet to be translated into service industry (see Table 4a).

Reflecting Korean industrial situation where big automobile and electronics companies heavily invest in new automobile technology, there were much interests in the usage of “automobile” devices along with newly attachable “parts.” “Brand” was emphasized to increase the market value of new products, especially linked to battery and wearable (Table 4b). In regard to social activities, “expert” showed the highest PBS score among the 30 keywords, directly linked to AI, big data and 3D printing (Table 4a). “Innovation” was also a salient theme to emcompass the network, and “certificate” reflected government's policy to nurture and authorize big data and 3D printing experts.

Fig. 6 demonstrates the reassembled network focusing on the concepts interlinking 13 technologies. The network represents the core structure of obligatory passage points that define the eco-system of emerging technologies. As immediately observed, artificial intelligence (AI) and robot are directly linked each other; and “expert” connects AI with big data. On the other hand, autonomous vehicle (AV), electrical vehicle (EV) and battery form a cluster that are linked to “system.” Other technologies like fintech, big data, IoT and wearable are closely related to “service.”

Regarding the network analogous to the Windrose model discussed in the methodology section, the overall representation poses the researcher an interpretive challenge to forecast the possible, and desirable, movement in the future. From the network of blog discourses, three pillars stand out as stark contrasts to the interest of on-line newspaper and government policy then. Firstly, despite the neglect of autonomous and electric vehicle industry by official media and existing industry, the bloggers’ representation has shown that those industries entered into a serious public consideration by forming a significant cluster of discourse. Secondly, the potential significance of AI and robots are manifested in the network while they had been overlooked in the official media representation during 2015 in Korea. On the other hand, the mention of AI and robots were usually related to sci-fi movies – with very few considerations of concrete technological trend. Panels concluded that this was partly due to the lack of indigenous industry that can compete with forerunners in Japan, US and UK. Thirdly, while the expertise, system development, and the creation of new service industry, which could be considered as “cloud seeds” of innovation, turn out to be significant in the network, they had not been met with adequate policy concerns. Spontaneous and effective movements to bring together these key variables are yet to be noticed. In this sense, these three concepts stand as a “constraint of reality” (Bauer & Gaskell, 2008) to achieve further development.

Table 4
Linked technologies with key concepts.

a. Linked technologies with expert, system, and service

Interlinked concepts	Technologies
Expert	AI
	3D printing
	Big data
Service	AV
	Fintech
	Big data
	IoT
	Wearable
System	AI
	AV
	EV
	VR
	IoT
	Healthcare
	Robot

b. Linked technologies with other concepts

Interlinked concepts	Technologies	Interlinked concepts	Technologies
Innovation	Big data	**Automobile**	Battery
	IoT		Robot
	VR	**Hardware**	IoT
	Battery		3D printing
Software	Fintech	**Charge**	AV
	3D printing		Battery
	IoT		EV
	Wearable		Wearable
Video	Big data	**Control**	Robot
	Drone		AV
Parts	AV		IoT
	EV	**Brand**	Battery
	Battery		Wearable
	3D printing	**Data**	IoT
Solution	IoT		Healthcare
	Big data		Wearable
	Fintech		
Certificate	Big data		
	3D printing		

4. Summary and discussion

While the analysis of mass media data in relation to scientific trajectory has precedence in academic studies (Bauer & Petkova, 2005; Hellsten, Dawson, & Leydesdorff, 2010; Kim, 2011), we tried to study the authenticity of media representation by employing alternative methods. Utilization of acceleration and skewness indices proposes a way to reveal important facets beyond the frequency of salience (Bauer & Petkova, 2005). In the Korean case, newspaper's representations of a few technologies, notably wearable, robot and healthcare, showed relatively high skewness and low acceleration rate despite their high media coverage. For the wearable, this was mainly due to the concentrated media coverage of Korean firms' activities when there were commercial events like the product release. For robot, it was mostly because of movie reviews that had robot-related themes from a philosophical perspective while any concrete discussion of technology was scarce. Healthcare technology was also discussed in a short period when there was a high expectation of financial market. In short, the mass media representations were barely spontaneous. More often than not, they reflected external interests apart from intrinsic technological developments.

From social representation theory (Moscovici, 2000) perspective, lay people's utterances and reflections are expected to complement the discovered limitation of official media representation when considering future changes. While some of individual expressions of blogs might look personal and even superficial, the emerged discursive structure highlights hidden values and concepts that might deserve more serious reflection beyond scattered emphasis of technology. For example, the on-going convergence of AI, robot and big data in the integrated semantic network is noticeable when official media had paid little attention at the time. Finally, the role of expert and service are manifested as importantly as the conventional interest of system development.

However, the extracted public discourses did not adequately represent technological characteristics of artificial intelligence or robots either. In this regard, the “Alpha-Go shock” (Korea Times, 10 March 2016) only reflected the public's unpreparedness to think about concrete facets of those technologies and their long-term impacts. The government's typical response after the Alpha-Go event also demonstrated that South Korea's nation technological strategy has been locked in by a top-down directive to follow the foreign

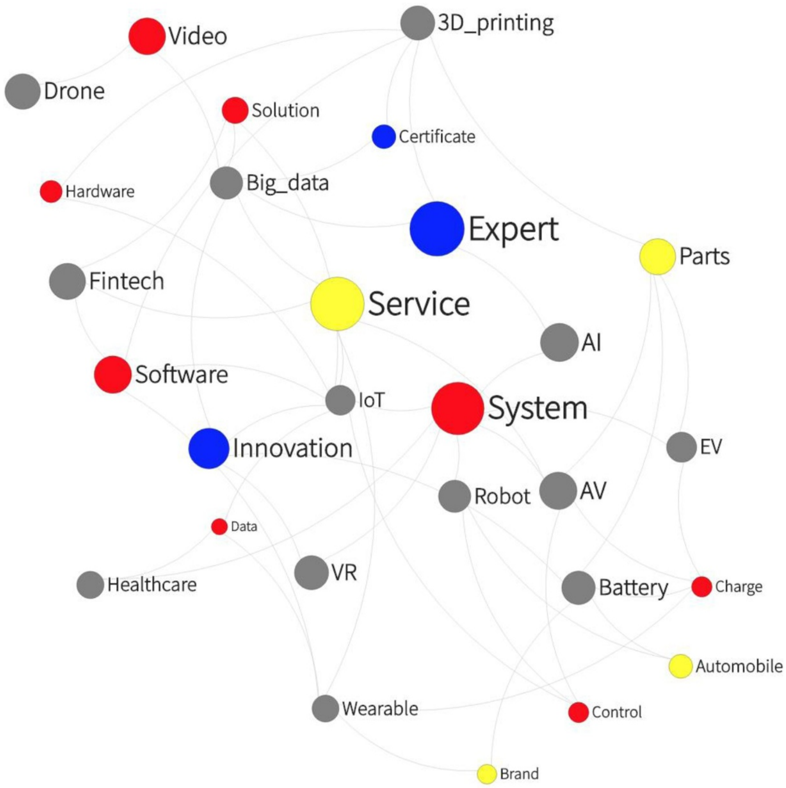


Fig. 6. Core network of technological concepts (node size: PBS of original network/red: tech yellow: industry blue: society).

precedence and invest in manufacturing field to produce (copy) a similar artificial intelligence system with significantly less cost. Even though this kind of governmental inertia is often met with severe criticism (Digital Daily, 19 December 2017), absence of alternative thoughts from the public might aggravate the situation.

Some science policy researches argue (for example, see Kim & Jang, 2018; Kim and Park, 2015) that the national innovation in South Korea is being hampered by the bureaucratic approach that is excessively dependent on formalized measures to justify decision making such as selection of promising technologies in the future while making little efforts to understand more subtle dynamics in the innovative environment. In contrast to such approach, the authors tried to derive key concepts that might incur some social reflections of technological trajectory beyond providing grounds for identifying emerging technologies. For instance, our research suggests that nurturing innate field experts and creating novel type of services, related to the technologies in question, deserves more significance. Still, our research might also bear similar limitation.

Two things should be pointed out for further research with theoretical considerations: Firstly, our reformulation of data output implies that the context in which a society constructs its unique characteristics of science and technology (Jasanoff, 2006) should be brought into policy considerations more actively. Questioning why spontaneous growth of experts and services in South Korean technological field is facing serious challenges (Financial Times, 19 August 2018) could enrich further discussions. Secondly, the comparison between newspaper and blog casts interesting points of social reflections. While it is evident that mass media do not necessarily stand in impartial position, there is little evidence that lay people's discourses alone could supplement the bias of the elite representation. As witnessed in South Korean blogs' representation of AI and robots, public imagination and discussions could lag behind what is actually happening in the industrial field. Although it is beyond our study, more active and spontaneous participation of field experts in communicative platforms and the interaction of heterogeneous media are expected to improve the quality of on-line discussions.

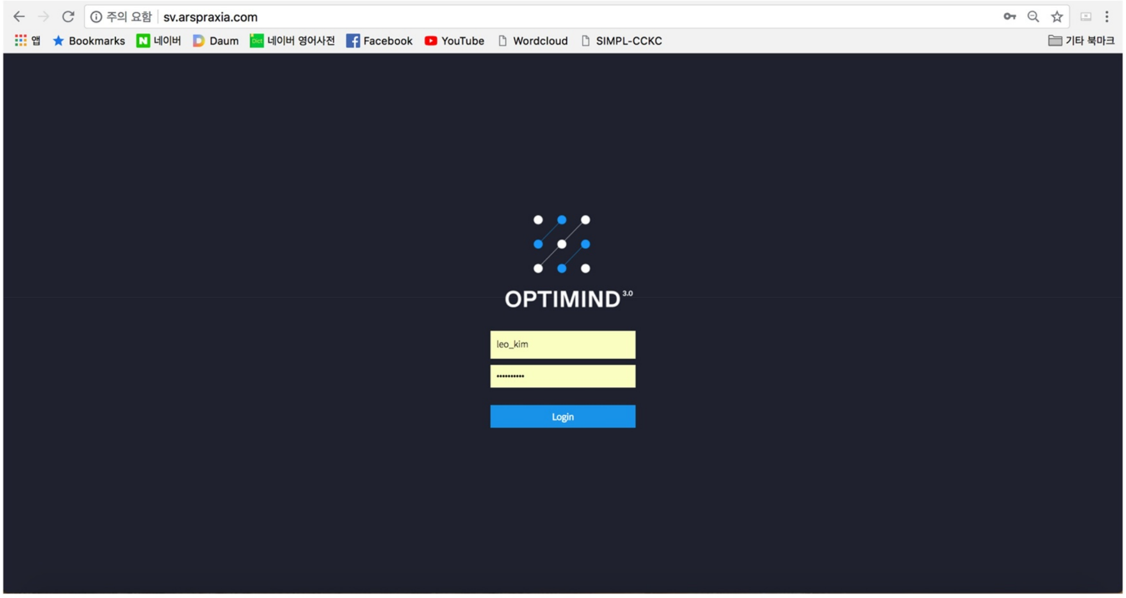
In this regard, sociological or social-psychological perspective and modeling could enhance the pathway of information processing. The authors have appropriated social-psychological theory to establish a research model to process unstructured data and utilize them to forecast technological movements related to social environment. For the future research, a comparative experiment to analyze the contents that represent higher level of expertise is expected to complement our study, improving ways to interpret digitized discourses and forecast social trends.

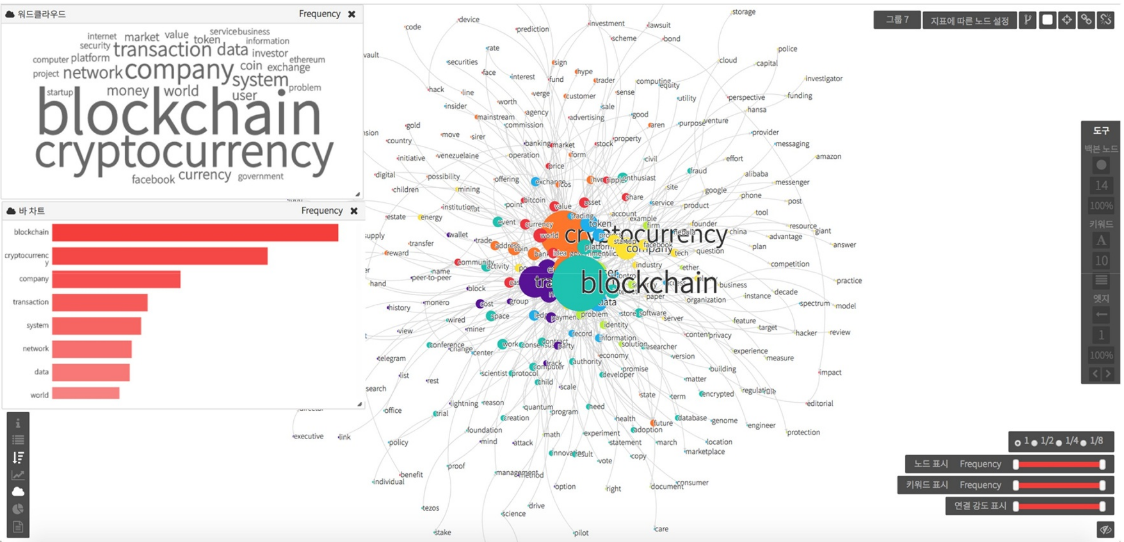
Acknowledgment

The authors are grateful to Byeongsoo Yu who provided invaluable advice for the use of statistical measures and significantly contributed to developing mathematical index. We are also thankful to anonymous reviewers who offered kind advices to improve this article.

Appendix I

The visual description of Optimind (ver. 2.0)

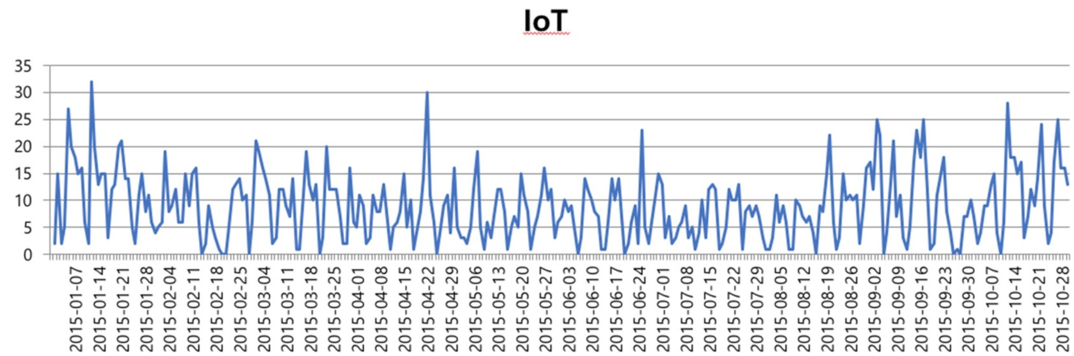




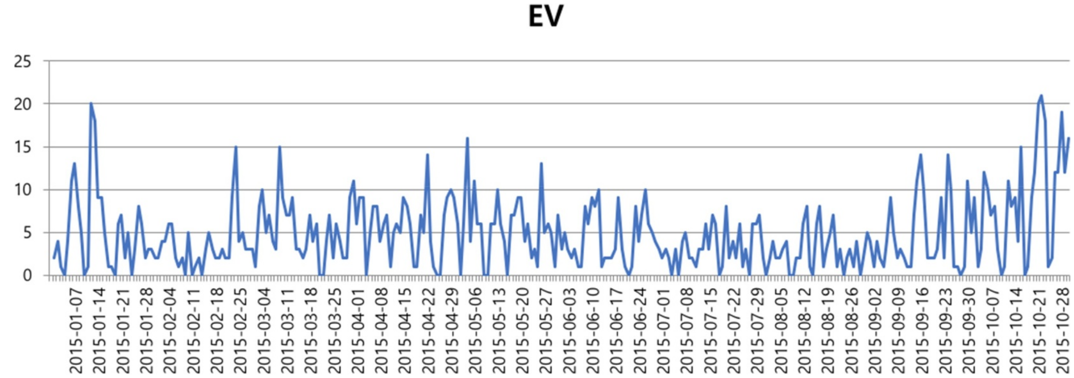
Appendix II

Daily news postings of each technology

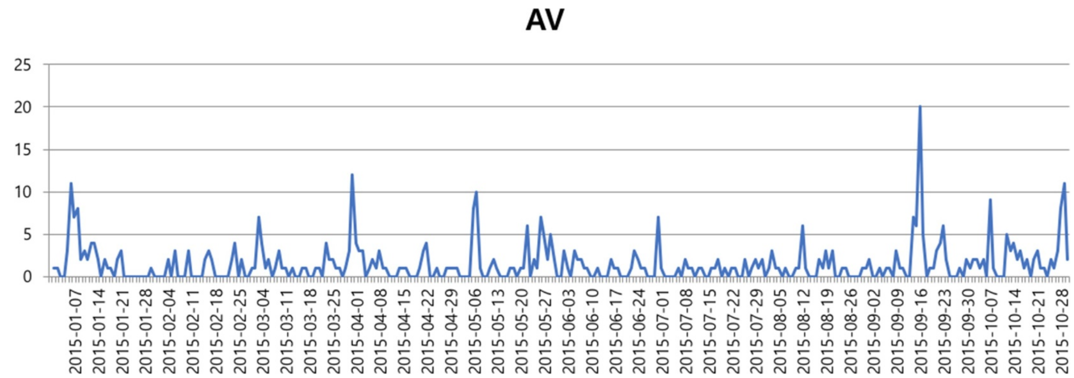
1. IoT



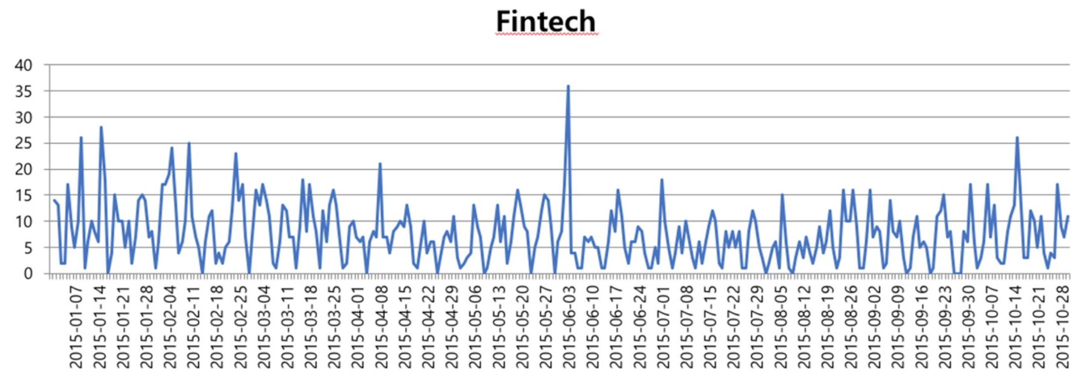
2. Electric vehicle



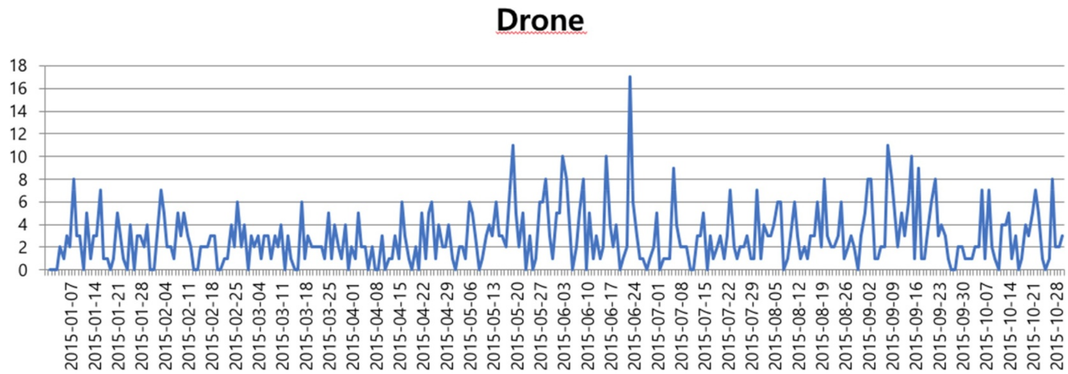
3. Autonomous vehicle



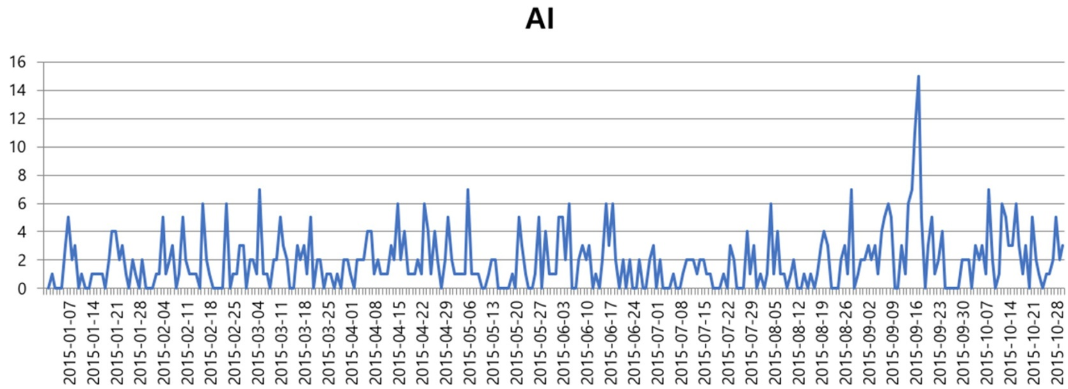
4. Fintech



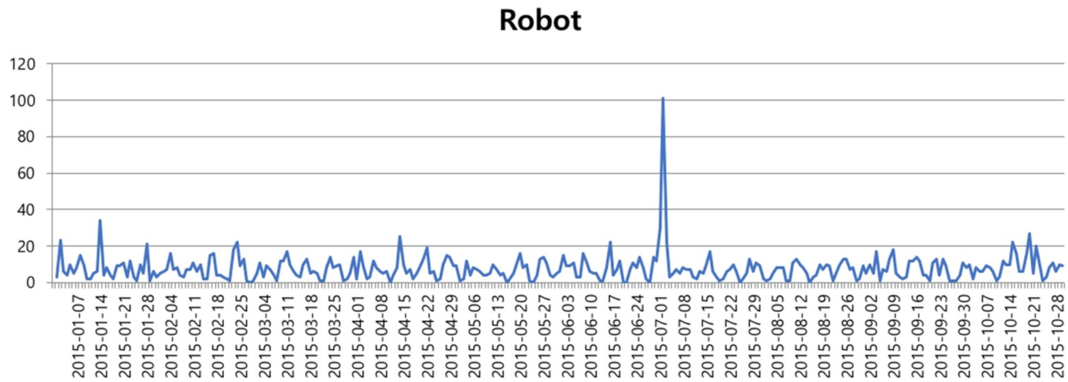
5. Drone



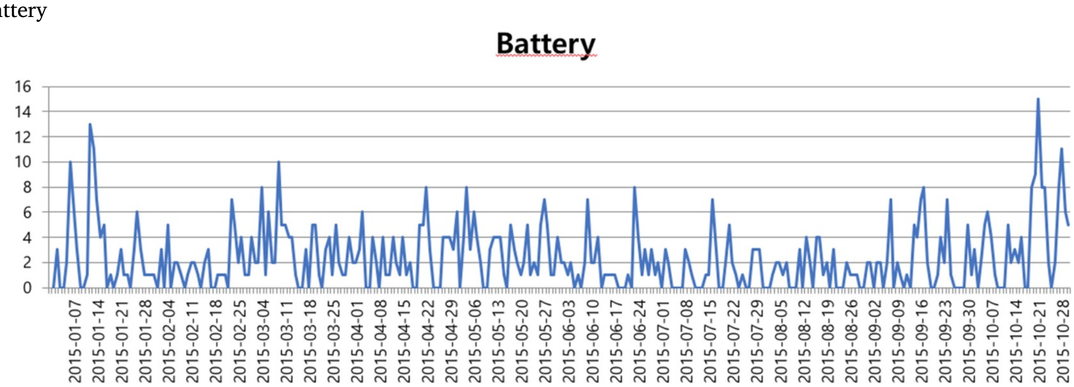
6. Artificial intelligence



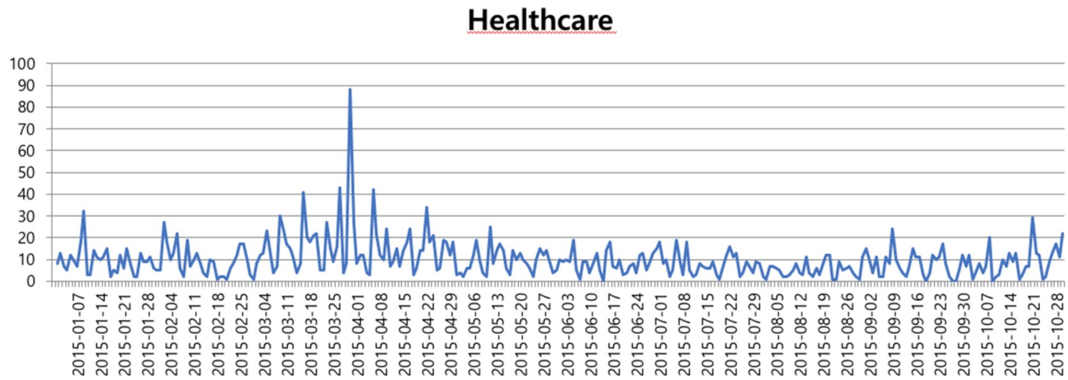
7. Robot



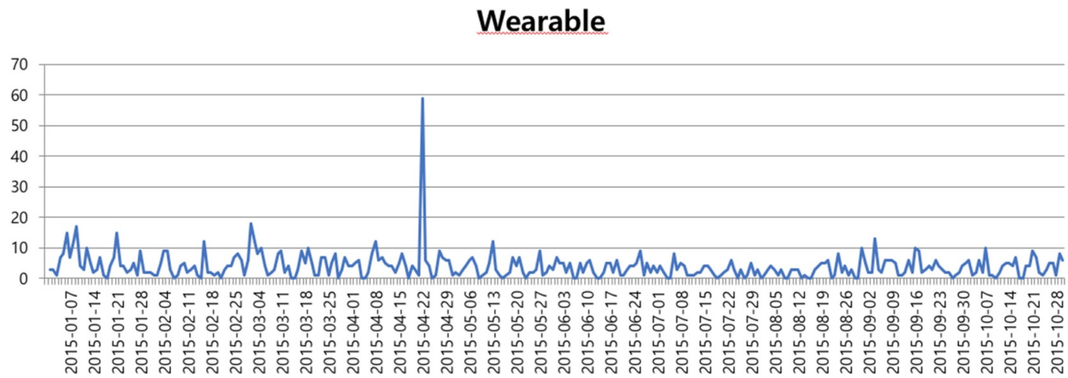
8. Battery



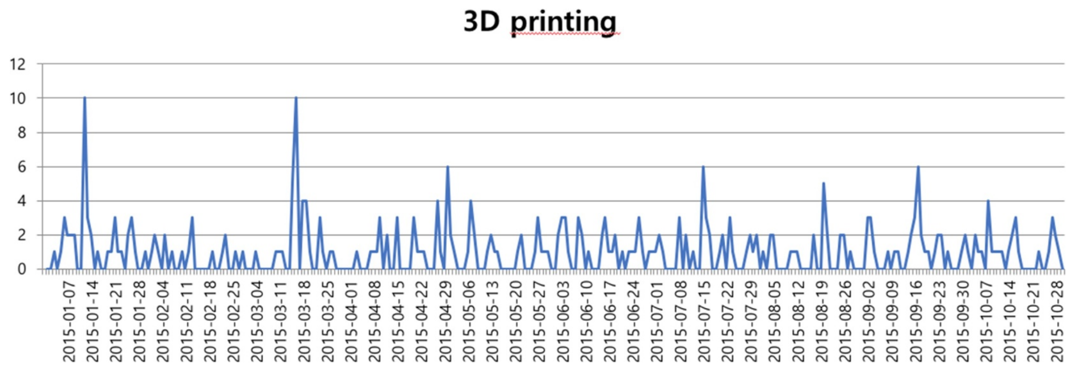
9. Healthcare



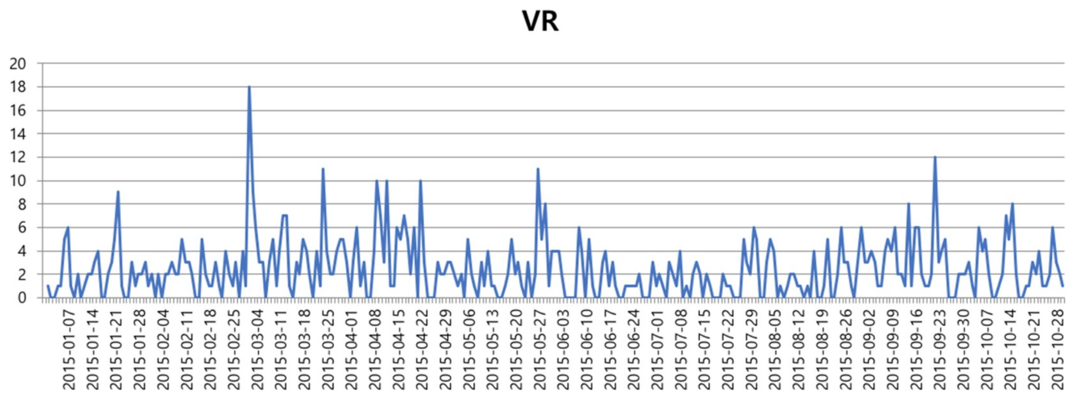
10. Wearable



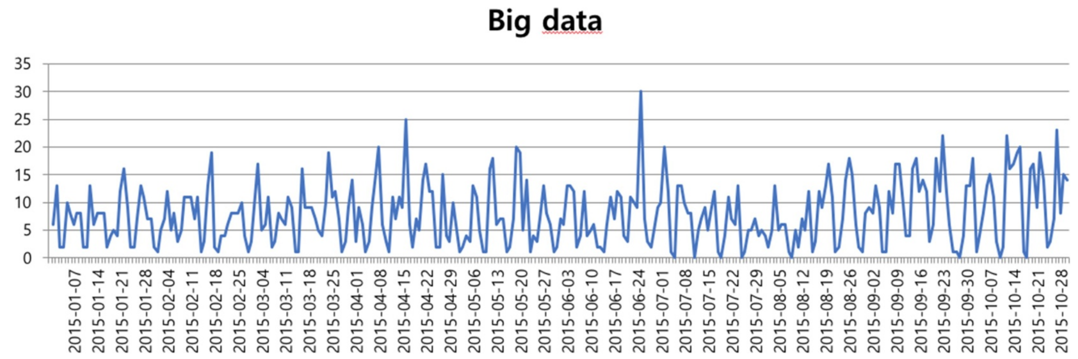
11. 3D printing



12. Virtual reality



13. Big data



Appendix III

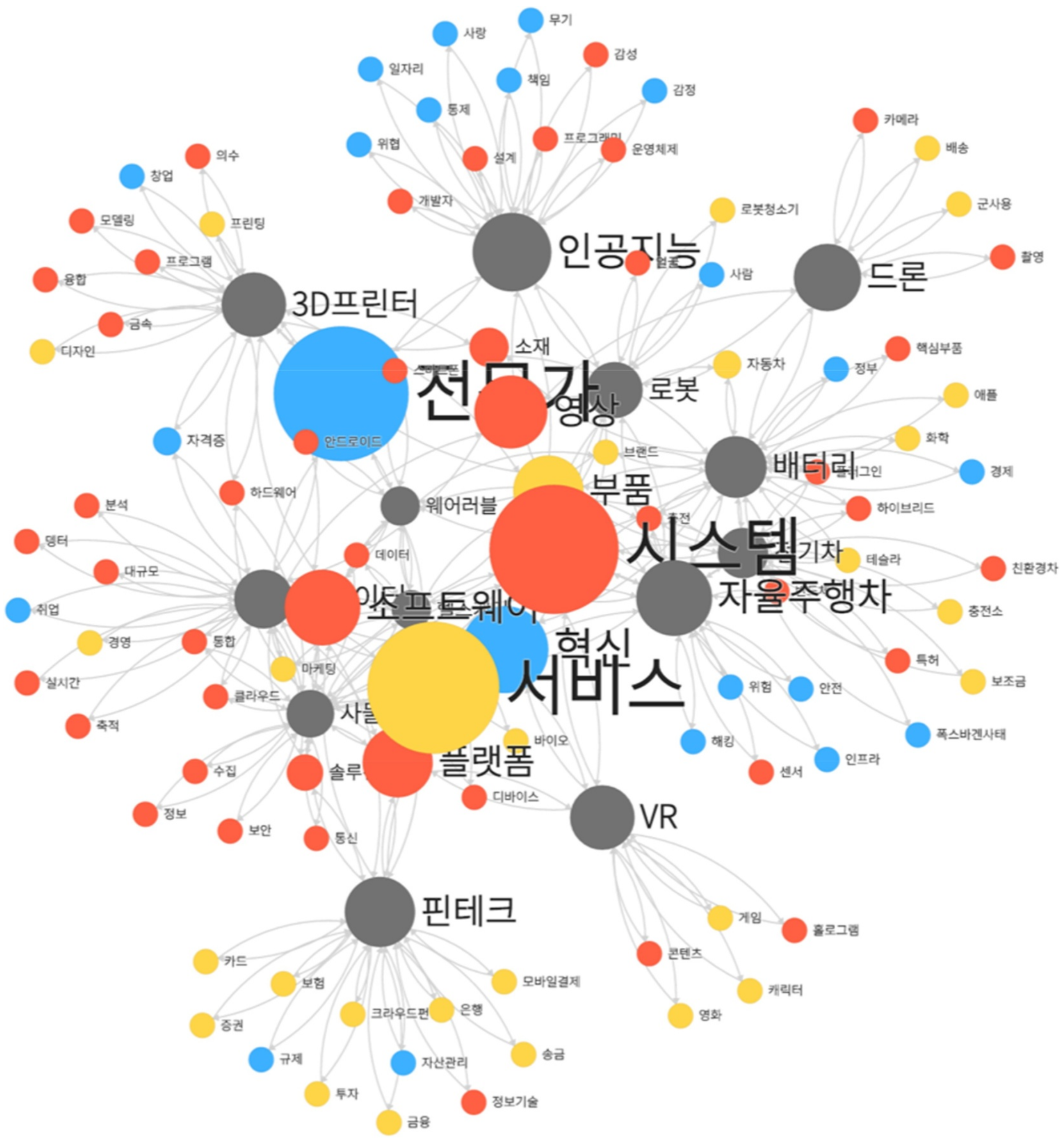
Statistical outputs of news media

	IoT	EV	AV	Fintech	Drone	AI	Robot	Battery	Healthcare	Wearable	3Dprinter	VR	Big data
Average	8.90759	4.87459	1.49175	7.60396	2.83828	1.84818	7.75248	2.36304	9.86799	3.98680	0.99010	2.43894	7.84818
Acceleration	0.00040	0.00020	0.00012	0.00010	−0.00004	−0.00004	0.00004	0.00014	−0.03860	0.00010	0.00001	0.00001	0.00016
Skewness	0.78151	1.25619	3.42265	1.16860	1.48146	1.93795	6.38574	1.55443	3.50394	6.51183	2.71078	1.84296	0.75388
Std. Deviation	6.27763	4.16117	2.28711	5.65231	2.48110	2.00745	7.67822	2.55539	8.45884	4.49501	1.39176	2.49942	5.55079

Appendix IV

The original picture of integrated network

- Grey: 13 technology labels
- Red: Technology related keywords
- Yellow: Business related keywords
- Blue: Keywords related to social activities
- The size of node represents potential boundary spanner measure



References

Barthes, R. (1967). *Elements of semiology, trans. Lavers a and smith c.* New York: Jonathan Cape Ltd.

Bauer, M., & Gaskell, G. (2008). Social Representations theory: a progressive research programme for social psychology. *Journal for the Theory of Social Behaviour, 38*(4), 335–353.

Bauer, M., & Petkova, K. (2005). Long-Term trends in the public representation of science across the 'Iron-Curtain': 1946-95. *Social Studies of Science, 34*(1), 99–131.

Borgatti, S., & Everett, M. (1997). Network analysis of 2-mode data. *Social Networks, 2*(3), 243–269.

Callon, M., Law, J., & Rip, A. (1986). *Mapping the dynamics of science and Technology: sociology of science in the real world.* London: Macmillan.

Carley, K. (1993). Coding choices for textual analysis: a comparison of content analysis and map analysis. *Sociological Methodology, 23*, 75–126.

Deaux, K., & Philogène, G. (2001). *Representations of the social.* Oxford: Blackwell.

Entman, R. M. (1993). Framing: toward clarification of a fractured paradigm. *Journal of Communication, 43*, 51–58.

Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., et al. (1998). Text mining at the term level. *Principles of Data Mining and Knowledge Discovery*, 65–73.

Freeman, C. (1979). Centrality in networks: conceptual clarification. *Social Networks, 1*, 215–239.

Gamson, W. A., & Modigliani, A. (1987). The changing culture of affirmative action. In G. Gaskell, & M. Bauer (Eds.). *(2001) biotechnology 1996–2000: the years of controversy*. London: Science Museum.

Georghiou, L. (Ed.). (2008). *The handbook of technology foresight: concepts and practice.* Edward Elgar Publishing.

Glenn, Jerome C., & Theodore, J. Gordon (Eds.). (2003). *Future research methodology*American Council for the United Nations University.

Hellsten, I., Dawson, J., & Leydesdorff, L. (2010). Implicit media frames: Automated analysis of public debate on artificial sweeteners. *Public Understanding of Science, 19*(5), 590–608.

Hogg, R. V., & Craig, A. T. (1995). *Introduction to mathematical statistics (5th ed.)*. Upper saddle river, New Jersey: Prentice Hall.
Jasanoff, S (Ed.). (2006). *States of Knowledge: the Co-production of science and the social order*. London: Routledge.
Jurowetzki, R., & Hain, D. S. (2014). Mapping the (R-)evolution of technological fields–a semantic network approach. In LM. Aiello, & D. McFarland (Vol. Eds.), *Social informatics: 8851*Springer, Cham. SocInfo 2014. Lecture Notes in Computer Science.
Kim, L., & Jang, D. (2018). Expert Views on Innovation and Bureaucratization of Science: Analysis on Discourses on Scientific Governance. *Science and Public Policy.*.
Kim, L., & Kim, J. (2018). Expert views on innovation and bureaucratization of science: semantic network analysis of discourses on scientific governance. *Science and Public Policy, 45*(1), 36–44.
Kim, L., & Kim, N. (2015). Connecting opinion, belief and value: semantic network analysis of a UK public survey on embryonic stem cell research. *Journal of Science Communication, 14*(1), A01.
Kim, J., & Lee, J. (2015). A Semantic Network analysis of technological innovation in dentistry: a case of CAD/CAM. *Asian Journal of Technology Innovation, 23*(1), 40–57.
Kim, L. and Park, H. (2015) Diagnosing "Collaborative Culture" of Biomedical Science in South Korea: Misoriented Knowledge, Competition and Failing Collaboration. East Asian Science, Technology and Society.
Kim, L. (2013). Denotation and connotation in public representation: semantic network analysis of Hwang supporters' internet dialogues. *Public Understanding of Science, 22*(3), 335–350.
Kim, L. (2011). Media framing of stem cell research: a cross-national analysis of political representation of science between the UK and South Korea. *Journal of Science Communication, 10*(3).
Kwon, K., Barnett, G., & Chen, H. (2009). Assessing cultural differences in translations: a semantic network analysis of the universal declaration of human rights. *Journal of International and Intercultural Communication, 2*(2), 107–138.
Latour, B. (2005). *Reassembling the social.* Oxford: Oxford University Press.
Lee, S., Lee, S., Kim, K., & Park, J. (2012). Bursty event detection from text streams for disaster management. *Proceedings of the 21st international conference on world wide web* (pp. 679–682). ACM.
Lent, B., Agrawal, R., & Srikant, R. (1997,. August) Discovering trends in text databases. In *KDD* (Vol. 97, pp. 227-230).
Li, N., & Wu, D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems, 48*(2), 354–368.
Long, J. C., Cunningham, F. C., & Braithwaite, J. (2013). Bridges, brokers and boundary spanners in collaborative networks: a systematic review. *BMC Health Services Research, 13*(158).
Lucini, F. R., Fogliatto, F. S., da Silveira, G. J., Neyeloff, J., Anzanello, M. J., Kuchenbecker, R. D. S., et al. (2017). Text mining approach to predict hospital admissions using early medical records from the emergency department. *International Journal of Medical Informatics,* 1–8.
Moscovici, S. (2000). *Social Representations: explorations in social psychology.* Cambridge: Polity Press.
Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: a systematic review. *Expert Systems with Applications, 41*(16), 7653–7670.
OECD. (2017). *Digital economy outlook 2017*. Paris: OECD Publishing.
Rickman, T., & Cosenza, R. (2013). The changing digital dynamics of multichannel marketing: the feasibility of the weblog: text mining approach for fast fashion trending. *Journal of Fashion Marketing and Management, 11*(4), 604–621.
Serrano, M. A., Boguñá, M., & Vespignani, A. (2009). Extracting the Multiscale Backbone of Complex Weighted Networks. *Proceedings of the National Academy of Sciences of the USA, 106*(16), 6483–6488.
Suerdem, A. (2013). A network based semiotic analysis: critical commentary on the 'field and dynamic nature of sense-making: theoretical and methodological implications'. *Papers on Social Representations, 22*(23), 1–13.
Sun, A., Lachanski, M., & Fabozzi, F. J. (2016). Trade the tweet: social media text mining and sparse matrix factorization for stock market prediction. *International Review of Financial Analysis, 48*, 272–281.
Vecchio, P. D., Mele, G., Ndou, V., & Secundo, G. (2018). Creating value from social big data: implications for smart tourism destinations. *Information Processing and Management, 54*(5), 847–860.