JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 19, No. 3, pp. 529-535, Mar. 2018

TF-IDF와 네트워크분석을 이용한 특허 데이터 분석과 경쟁우위 전략수립에 관한 연구

윤 석 용[1] · 한 경 석[2*]
[1]숭실대학교 IT정책경영학과 박사수료
[2]숭실대학교 경영학부 경영정보시스템 교수

A Study on Patent Data Analysis and Competitive Advantage Strategy using TF-IDF and Network Analysis

Seok-Yong Yun[1] · Kyeong-Seok Han[2*]
[1]Dept. of IT Policy Management, Graduate School, Soong-sil University, Seoul 06978, Korea.
[2]*Professor of MIS, School of Business Administration, Soong-sil University, Seoul 06978, Korea.

[요 약]

데이터는 폭발적으로 증가하고 있으나 아직도 많은 기업이 데이터 분석을 현황 설명(descriptive analysis)이나 진단 분석(diagnostic analysis)에만 활용하고 예측분석(predictive analysis)이나 기업의 기술전략 분석 등에는 적절하게 활용하고 있지 못하다.

본 연구는 오픈 되어 있는 특허의 IPC 코드, 발명자, 출원일 등의 정형데이터와 청구항 등의 비정형 데이터를 네트워크분석, TF-IDF 등의 빅데이터 분석기법을 활용하여 경쟁기업의 확보 기술과 핵심 기술의 분포, 해외 진출 전략을 파악하기 위한 분석 프로세스를 제시하고 이를 데이터 분석을 통하여 증명하고자 한다.

[Abstract]

Data is explosively growing, but many companies are still using data analysis only for descriptive analysis or diagnostic analysis, and not appropriately for predictive analysis or enterprise technology strategy analysis.

In this study, we analyze the structured & unstructured patent data such as IPC code, inventor, filing date and so on by using big data analysis techniques such as network analysis and TF-IDF. Through this analysis, we propose analysis process to understand the core technology and technology distribution of competitors and prove it through data analysis.

색인어 : 오픈소스, 네트워크 분석, TF-IDF, 빅데이터, 예측분석, 특허 데이터 분석

Key word : Open Source, Network Analysis, TF-IDF, Big Data, Predictive Analysis, Patent Data Analysis

http://dx.doi.org/10.9728/dcs.2018.19.3.529

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-CommercialLicense(http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 22 March 2018; **Revised** 27 March 2018
Accepted 29 March 2018

***Corresponding Author; Kyeong-Seok Han**

Tel: +82-10-8954-7137
E-mail: kshan@ssu.ac.kr

Ⅰ. 서 론

10여년 전만 하더라도 데이터는 정보(information)와 지식(knowledge)을 만들기 위한 원천 데이터(raw data)의 역할만을 요구할 뿐 그 가치에 대한 평가를 제대로 받지 못했다. 영국의 이코노미스트지는 최근호에 "세계에서 가장 가치 있는 자원은 석유가 아닌 데이터" 라는 조금은 진부한 글을 실었다. 그러나 여기서 주목할 내용은 부제에 있는 "반독점(antitrust)" 으로, 이제 데이터는 가치의 중요성을 넘어 글로벌 통제가 되지 않는다면 세계 경제에 위험이 될 수도 있다는 이슈를 제기하고 있다[15].

데이터의 증가속도는 세계적으로 2011년 약 2제타 바이트의 데이터가 2016년 8제타 바이트로 5년 만에 4배 이상 증가하고 있고 있음을 보여주고 있다[10].

이는 데이터의 폭발이라고 할 만하며 IoT(internet of things)의 세상이 됨으로써 그 증가 속도가 기하급수적이다. 그러나 많은 기업에서 아직도 데이터 분석을 현상 설명이나 진단분석(descriptive & diagnostic analysis)에 머물러 있을 뿐 예측분석(predictive analysis)나 기업전략수립에 적절하게 활용하지 못하고 있다[1].

본 논문에서는 빅데이터 분석의 다양한 통계적 기법과 기계학습(machine learning)을 이용하여 기업의 경쟁우위 전략을 확보하기 위한 실용적 방법을 제시하고자 한다.

경쟁기업에 대한 우위 전략을 수립하기 위해서는 경쟁사를 분석하는 것이 무엇보다 중요한데 이는 데이터 즉 팩트(fact)에 기반 해야 함은 당연하다. 그러나 경쟁사의 데이터는 쉽게 획득이 어렵고, 상장되어 있는 기업이 아니면 기업 재무 데이터에 대한 접근조차도 쉽지 않다. 설령 기업 재무 데이터를 확보했다 하여도 이는 기업의 과거에 대한 재무적 정보만을 알려줄 뿐 기업의 중·장기적인 방향, 더 나아가 핵심 기술 전략을 파악하기는 불가능 하다[3, 4].

그러나 기업은 개발된 기술의 배타적 권리를 확보하기 위하여 특허 등록을 하고, 등록된 특허는 공개되어 20년간 법적 보호를 받게 된다[5].

본 논문은 이러한 공개된 특허관련 정형 데이터와 비정형 데이터를 이용하여 경쟁사의 기술전략을 분석하기 위한 데이터 분석방법 및 절차를 제시함으로써 경영우위 전략수립을 위한 실용적인 방법을 제시하고자 한다.

Ⅱ. 이론적 배경

2-1 특허 데이터

특허제도는 발명을 장려하고 국가산업의 발전을 도모하기 위하여 법으로 규정된 제도로써 발명 특허권을 받기 위해서는 산업 이용가능성과 기술의 신규성, 그리고 진보성을 갖고 있어야 한다[6].

이러한 특허권은 출원인이 심사를 청구하고, 1년 6개월이 경과되면 일반에게 특허를 공개함으로써 20년간 배타적으로 권리를 보호받게 된다[17].

2-1-1 특허 데이터 구성

특허 데이터는 KIPRIS(korea intellectual property rights information service)를 통하여 출원번호, 발명의 명칭, IPC(international patent classification) 코드, 출원인, 발명자, 출원일자, 등록일자, 공개일자, 국제출원번호, 국제출원일자, 패밀리정보 등의 정형 데이터와 초록, 청구항, 도면 등의 비정형 데이터를 제공하고 있다[7, 18].

2-1-2 국제 특허 분류의 구조

국제특허분류코드인 IPC 코드는 1968년 제정되었고 현재 7만여개의 코드로 구성되어 있으며 백여개의 회원국이 공유하고 있다. Fig. 1과 같이 섹션, 클래스, 서브클래스, 메인그룹, 서브그룹으로 구성되어 있고, Table 1과 같이 각 섹션은 기술 분류별로 구성되어 있다. 이러한 IPC 코드는 각국의 기술 분류의 세분화에 적정하지 못한 면도 있어 독자적인 코드를 사용하기도 하지만 많은 나라에서 통용되고 있어 국제 특허분류 표준코드의 역할을 하고 있다[14].

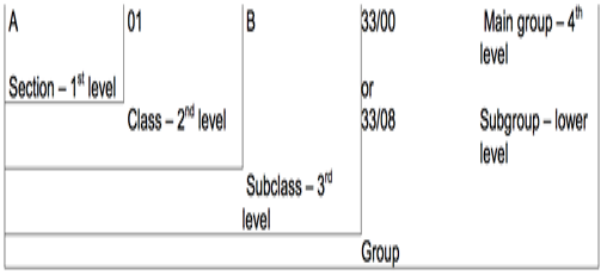

그림 **1.** IPC 코드 체계
Fig. 1. hierarchical structure of IPC

표 **1.** IPC 섹션 분류
Table 1. section classification of IPC

Section	A HN	B POT	C CM	D TP	E FC	F MEL	G PHY	H ELE
IPC	8,498	16,778	14,449	3,050	3,250	8,551	8,011	8,283

Table 1에서 human necessities는 HN으로 performing operations transporting는 POT로 chemistry metallurgy는 CM로 textiles paper은 TP로 fixed constructions는 FC로 mechanical engineering lighting은 MEL로 physics는 Phy로 electricity는 ELE로 표기하였다.

2-2 네트워크 분석

2-2-1 네트워크 중심성 이론

네트워크(network)로 연결되어 있는 영향력 있는 노드(node)에 대한 중요성을 파악하기 위한 이론으로써 연결중심성, 매개중심성, 인접중심성 지표가 있다[9].

특허 데이터의 IPC 코드를 이용한 네트워크 분석을 통하여 경쟁 사의 특허 전략과 핵심 기술을 파악할 수 있다.

2-2-2 연결중심성(degree centrality)

노드의 중요도를 나타내는 지표로, 노드가 가지는 연결의 수를 나타낸다. 방향성의 엣지(edge)가 있는 경우 입력연결중심성(in-degree centrality)과 출력연결중심성(out-degree centrality)로 나뉘어 입력과 출력에 대한 연결의 영향도를 평가할 수도 있다.

필요에 따라 가중 엣지(weighted edge)를 이용하여 입력과 출력의 영향도를 평가하기도 하며 연결중심성이 크면 네트워크에서 핵심적인 역할을 담당하는 노드가 된다.

2-2-3 매개중심성(betweenness centrality)

어떤 노드에 대하여 그 노드를 통해 연결되는 노드의 수로, 큰 군집을 연결하는 노드는 매개중심성이 상대적으로 높다고 할 수 있다.

매개중심성이 큰 노드가 사라질 경우 군집간의 연결이 사라지므로 네트워크에서 중요한 위치를 갖게 된다.

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (1)$$

$\sigma_{st}(v)$ = Number of paths through node v in the shortest path
σ_{st} = Total number of shortest path from node s to nodes t

2-2-4 인접중심성(closeness centrality)

네트워크에서 특정 노드가 다른 노드들과 가까이 있는 정도를 나타내는 지표로써 노드간 거리가 짧다면 접근성과 전달성이 좋다고 할 수 있다.

$$C_C(v) = \frac{1}{\sum_{i \neq v} d(v,i)} \quad (2)$$

$d(v,i)$ = Distance from node v to node i

2-2-5 TF-IDF를 이용한 자료 자동분류

TF-IDF(term frequency - inverse document frequency)는 정보검색과 텍스트 마이닝에서 이용하는 가중치 계산방법으로 이를 이용하여 특허 자동분류에 응용이 가능하다.

TF-IDF는 한스 피터 룬이 제안한 정보 검색 자동화를 위한 이론적 배경으로 활용되고 있으며 계산 방식은 다음과 같다.

$$w_{t,d} = tf_{t,d} \times \log\left(\frac{N}{df_t}\right) \quad (3)$$

$tf_{t,d}$ = frequency of term in documents
df_t = number of documents containing term
N = total number of documents

$$tf_{t,d} = 0.5 + \frac{0.5 \times f_{t,d}}{\max\{f_{w,d} : w \in d\}} \quad (4)$$

$$idf_{t,d} = \log \frac{|D|}{1 + |\{d \in D : t \in d\}|} \quad (5)$$

Ⅲ. 경쟁사 특허분석을 통한 경쟁우위 전략 수립

본 논문에서는 철강업계의 P사의 가상의 경쟁사인 H사의 특허를 분석하기 위하여 2000년부터 2014년까지 15년간 등록된 5311건의 특허 데이터를 KIPRIS를 통하여 확보하였다. 데이터는 IPC 코드, 특허 등록일, 발명자, 출원인, 특허 청구항 등 55개 항목으로 구성되어 있고 KIPRIS에서 제공되지 않는 정보는 한국과학기술정보연구원, 특허청, 구글 등의 검색을 통하여 수집하였다.

연구가설은 경쟁사의 등록된 특허 데이터를 분석하면 핵심기술 확보, 핵심기술 분포, 해외 진출전략을 파악하고 이를 기반으로 특허 자동 분류 체계를 구축할 수 있음을 아래 프로세스로 증명하고자 한다.

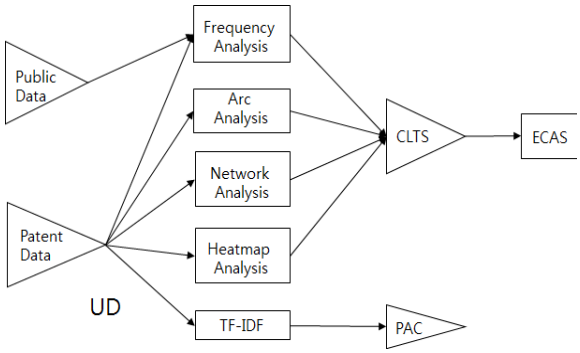

그림 **2.** 특허 데이터를 이용한 경쟁전략 모형
Fig. 2. competitive model using patent data analysis

Fig. 2와 같이 competitors long-term technology strategy는 CLTS로 patent auto classifier은 PAC로 establish competitive advantage strategy는 ECAS로 unstructured data는 UD로 표기하였다.

3-1 특허 데이터의 탐색적 분석

특허 데이터의 IPC 코드 전체를 분석 대상으로 하면 항목이

너무 세분화 되므로 IPC 코드의 섹션과 클레스로 코드를 재구성하여 등록된 특허의 빈도를 살펴보면 Fig. 3과 같이 특정한 코드로 집중되어 있음을 알 수 있다. 즉 철야금, 금속가공, 주편 관련 특허가 많지만 운송, 포장, 측정시험 등에서도 특허가 증가하고 있어 제품의 품질과 운송 등에 많은 노력을 하고 있음을 알 수 있다.

그리고 조강생산량의 증가와 주가의 변동이 특허 출원건수와 동기화되고 있음을 Fig. 4, Fig. 5를 통하여 파악할 수 있다.

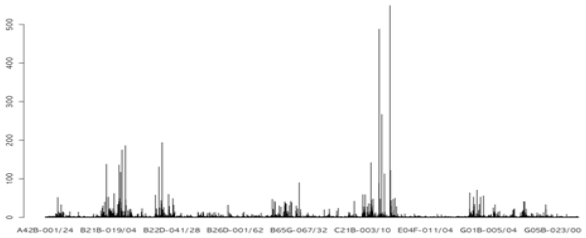

그림 **3.** IPC 코드별 빈도분석
Fig. 3. frequency analysis by IPC

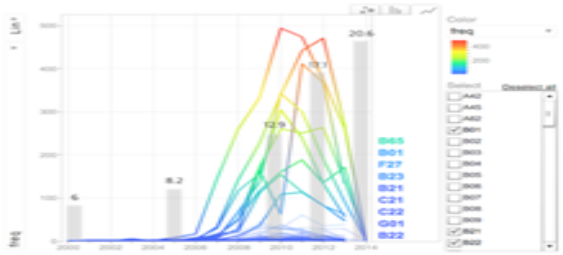

그림 **4.** 조강생산량과 특허 등록건수
Fig. 4. production amount & patent registration

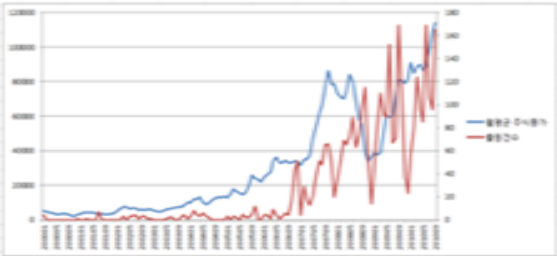

그림 **5.** 주가 변화와 특허 등록건수
Fig 5. stock prices & patent registration

또한 특허 패밀리 분석은 경쟁사의 해외전략을 읽을 수 있는 정보를 제공한다. 특허 패밀리는 국내 특허를 출원하면서 동시에 외국에 출원하는 것으로 써, 특정 국가에 특허 출원이 집중될 경우 경쟁사의 해외진출 전략을 간접적으로 파악 할 수 있다. H사는 특허의 국내 단독 출원 대비 특허 패밀리 출원 비율이 낮아 해외진출을 적극적으로 추진하고 하고 있다고 판단하기 어려우나 중국 등 특정 국가에 특허 출원이 집중되고 있음을 알 수 있다.

그림 **6.** 패밀리 특허 등록 현황
Fig. 6. Family Patent Registration

그리고 등록된 특허의 인용분석을 통하여 핵심 특허를 분석할 수 있다. H사의 경우 소결공정과 연료원료장치에 특허가 집중됨을 Fig. 7과 같이 파악되었으며 피인용 특허가 가장 높은 특허는 KR2009008*****와 KR2010011*****로 각각 15회와 13회 피인용되었다.

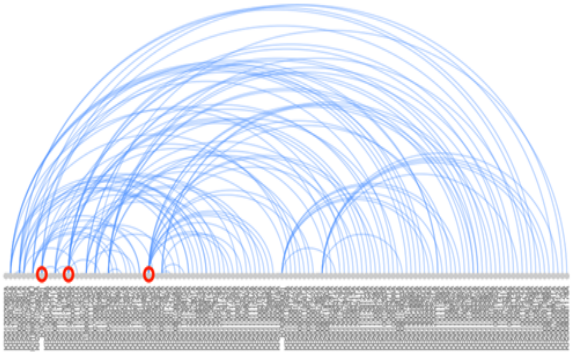

그림 **7.** IPC별 아크분석
Fig. 7. arc analysis of IPC

또한 등록된 특허의 상호 인용밀도(density)는 낮아 특허간 독립성이 높고 근접중심성에 분진제거장치관련 특허가 있은 것으로 보아 2010년 이후 환경관련 특허에 집중하고 있음을 알 수 있고, 중계중심성에 제강 및 냉연 관련 특허가 있어 자동차 강판제조에 주력하고 있음을 알 수 있다.

표 **2.** IPC를 이용한 네트워크 분석
Table 2. network analysis of IPC

N/W analysis	statistic	IPC code
closeness centrality	KR2009000***** : 0.0001 KR2010002***** : 0.0001	C23C-022/72 B03C-001/22
degree centrality	KR2009002***** : 8 KR2008010***** : 4 KR2012003***** : 3	B22D-011/16 C21D-008/02 B22D-011/124

H사의 15년간 등록된 전체 특허의 IPC 코드를 이용하여 히트맵(heatmap)을 그려보면 특정 IPC 코드가 특정 기간 동안 집중적으로 등록되어 기술력 확보에 노력하고 있음을 알 수 있고 특정 기술은 앞으로도 지속적으로 확보하려는 노력을 하고 있음을 알 수 있다.

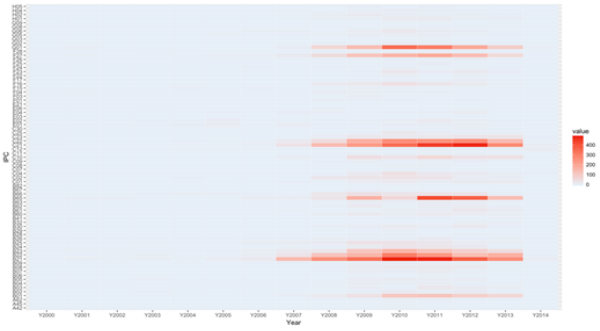

그림 **8.** IPC 히트맵 분석
Fig. 8. heatmap analysis of IPC

3-2 TF - IDF를 이용한 특허 자동분류

특허 IPC 코드와 경쟁사 정보를 이용하여 경쟁사의 공정별 분류맵을 구성하고 각 공정별 훈련용 데이터셋을 특허 데이터의 청구항에서 추출하여 TD(term-document)를 이용하여 공정별 주요 키워드(keyword)를 추출하여 Table 3과 같이 작성하였으며, MP는 mainprocess, SP는 subprocess, iron making는 IM, steel making는 SM,, rolling은 Roll, sintering는 Sin, subsidiary products는 Sub, blast furnace는 BF, refining는 RE, refinement, 는 RM, continuous casting는 CC, hot rolling, cold rolling는 각각 HR과 CR로 표기하였다.

표 **3.** 공정 분류 및 공정별 키워드
Table 3. process & keywords by process

MP	SP	code	keyword
IM	Sin	0101	iron ore subsidiary, raw material, sized ore, sinter limestone.
	Coke	0102	coke, bituminous coal, powdered coal, pulverized coal
	Sub	0103	COG, BOG, gas, crude light oil, cement
	BF	0104	pig iron, torpedo car
SM	RE	0201	iron, scrap, subsidiary raw material, oxygen, conventional blowing, argon, nitrogen.
	RM	0202	RH, degassing, inclusion, LF, desulfur, powder
	CC	0203	turndish, mold, cooling, slab, bloom, billet
Roll	HR	0301	reheating furnace, roughing mill, roll, cooler, winding, hot coil
	CR	0302	automobile, annealing, galvanize, plating, steel sheet, directional properties.

그리고 분류 대상 특허에서 단어를 추출하여 분류사전과 공정별 유사도를 계산하고 TF-IDF로 계산된 통계량을 이용하여 Table 4와 같이 분류할 수 있었으며 예측정확도(accuracy)는 76%를 나타냈다. 이는 향후 SVM(support vector machine) 등의 통계적 기계학습방법과 Word2Vec 등을 이용하면 업무에 적용할 수 있는 단계까지 올릴 수 있음을 확인할 수 있었다.

표 **4.** TF-IDF를 이용한 특허 분류
Table 4. patent classification using TF-IDF

No.		KR20120 06 ****	KR20110 06 ****	KR20110 01 ****	KR20100 10 ****	KR20090 07 ****	KR20090 10 ****
IM	Sin	100	36	0	0	0	0
	Coke	200	26	0	0	**1007**	106
	Sub	0	203	0	0	100	106
SM	BF	200	28	0	74	100	106
	RE	300	100	0	0	0	10
Roll	RM	**1365**	110	10	0	0	110
	CC	100	200	**987**	**1580**	0	300
	HR	200	**300**	0	100	0	**728**
	CR	0	100	0	0	0	100

Ⅳ. 결론 및 향후 연구 방향

본 논문은 공개되어 있는 특허 데이터와 빅데이터 분석기법을 이용하여 경쟁사의 기술전략을 파악하여 경쟁우위 전략을 수립하기 위한 실 사례와 방법을 제시하였다. 이를 다시 정리해 보면,

첫째, 경쟁사의 오픈 되어 있는 특허 데이터를 이용하여 frequency analysis, arc analysis, network analysis, heatmap analysis를 수행하게 되면 경쟁사의 중·장기적인 기술력 확보 및 해외 진출 전략을 파악할 수 있고 이를 활용하여 경쟁우위계획을 수립 할 수 있다.

둘째, 특허 데이터의 초록과 청구항의 비정형 데이터를 TF-IDF를 이용하여 분석하게 되면 특허를 자동 분류할 수 있게 되어 기술 분류 업무의 효율성을 확보할 수 있다.

특허는 기업이 개발한 기술의 배타적 권리를 법적으로 보호받기 위하여 특허청에 등록하고 일정기간 후 부터 오픈 된다. 따라서 경쟁사 뿐만 아니라 자사의 특허 데이터를 쉽게 확보할 수 있고 Fig. 2와 같은 분석 프로세스를 거치면 경쟁사에 대한 경쟁우위전략을 만들 수 있어 활용도가 넓다.

본 연구는 기술지도(Technology Map)의 작성이 IPC 코드로

만 구성되어 있어 경쟁사의 기술지도와는 차이가 있을 수 있어 많은 비정형 또는 정형 데이터를 이용하여 정교화 하는 방안을 모색하고 이를 기반으로 앙상블(ensemble), 딥러닝(deep learning) 등의 다양한 기계학습(Machine Learning) 방법을 이용한다면 보다 안정적인 모형 및 방법론 개발이 가능하다.

참고 문헌

[1] S.Y.Park and H.W.Park. "The Relationship between Technology Innovation and Firm Performance of Korean Companies based on Patent Analysis". Korea technology innovation society. Vol. 9, No. 1, pp. 1-25, 2006.

[2] J, B, Woo and Y, M, Jung, "Analysis of Factors Influencing Patent Citations", Korea Society for information management Vol. 27 No. 1, pp.103-118 2010.

[3] S. H. Jun "Technology Forecasting of intelligent systems using patent analysis", Korea intelligent information system society, Vol 21, No. 1, pp.100 -105, 2011.

[4] H. J. Kang, "Study on the Projection of the Promising Fusion Technologies by US Patent Analysis", KOSIME, Vol. 14 No. 3 pp. 111-115. 2007.

[5] S, H, Jun, S, S, Park,, " Forecasting Vacant Technology of Patent Analysis System using Self Organizing Map and Matrix Analysis", KOCON Vol. 10 No. 2, pp. 462-480, 2010.

[6] J, H, Choi, H, S, Kim " Keyword Network Analysis for Technology Forecasting", KIISS, Vol. 17 No. 4, pp.227-240 2011.

[7] C, H, Kim, "Identifying the linkage between technologies using co-classification analysis", KIOIE Vol. 2006, No. 11, pp. 711-717, 2006.

[8] G. J. Gwon, "The determinants of technology innovation in ICT firms", KIAA Vol. 28, pp. 139-157, 2009.

[9] W. J. Kim, "Network Centricity Theory", Koos Book, Vol. 111, No. 23, pp. 109-111, 2015.

[10] Ministry of Internal Affairs and Communications(2012), Japan, "Information Communication White Paper"

[11] Max H. Boisot, Ian C. MacMillan, Kyeong Seok Han. "Explorations in Information Space", Oxford University, Vol. 228, No. 24, 2007

[12] Jun Sunghae, "A Big Data Learning for Patent Analysis", Journal of Korean Institute of Intelligent System Vol23, No. 5, 2013.

[13] Gerard Salton, "Term-weighting approaches in automatic text retrieval", Information Processing & Management Vol24 No. 5 1988.

[14] "Guide to the International Patent Classification", WIPO, 2017

[15] The Economist, https://www.economist.com/news/leaders/21721656-data-economy-demands-new-approach-antitrust-rules-worlds-most-valuable-resource

[16] World Intellectual Property Organization, http://www.wipo.int

[17] Patent office, http://www.kipo.go.kr

[18] Patent Information Net, http://www.kipris.or.kr

윤석용(Seok-Yong Yun)

1987년 : 연세대학교 공학사
2000년 : 숭실대학교 정보과학대학원 석사
2017년 : 숭실대학교 대학원 IT정책경영학과 박사수료

1990년~1994년: ㈜포항제철 전산시스템 사원
1994년~1999년: ㈜현대정보기술 컨설팅 책임
2000년~2015년: ㈜포스코경영연구원/포스코 빅데이터 팀리더
2015년~현재: ㈜베가스 빅데이터 부사장
2016년~현재: 동국대학교 겸임교수
※관심분야 : 빅데이터분석, 분석방법론, 기계학습, 데이터 거버넌스, 개인정보보호 및 비식별화, 정보시스템감리 등.

한경석(Kyeong-Seok Han)

1983년 : 서울대학교 대학원 경영학과 (경영학석사)
1989년 : 미국 Purdue Univ, Krannert School of Management (경영학박사 - MIS)

1988년~1989년: 텍사스대학교(Austin) 연구원 (겸임)
1989년~1990년: 미국 휴스턴 대학교 조교수
1990년~1993년: 영남대학교 조교수
1997년~1999년: 한국경영기술 컨설턴트 협회 이사
1998년~2000년: 미국 펜실베니아대학교 와튼경영대학 교환교수
1993년~현 재: 숭실대학교 경영학부 교수
2015년~현 재: CMC Global(UN 산하 NGO인 Certified Management Consultants 기구) 부회장 (아시아지역 대표)
※관심분야 : Technical MIS, Digital Economy, e-Business, ERP, Agent_Based Simulation, 전자상거래, 중소기업정보화, 기업컨설팅, 기업자금지원정책 등