Journal of the Korea Convergence Society
Vol. 11. No. 7, pp. 217-222, 2020

ISSN 2233-4890 / eISSN 2713-6353
https://doi.org/10.15207/JKCS.2020.11.7.217

데이터 품질관리 평가 모델에 관한 연구

김형섭
한양대학교 일반대학원 경영컨설팅학과 박사과정

A study on the data quality management evaluation model

Hyung-Sub Kim
Doctoral Course, Hanyang University Student, Division of Management Consulting

요 약 본 연구는 데이터 품질관리 평가 모델에 관한 연구이다. 정보통신기술이 고도화되고 저장 및 관리에 대한 중요성이 증가를 하기 시작하면서 데이터에 대한 관심이 증가를 하고 있다. 특히 최근에는 4차산업혁명과 인공지능에 대해 관심이 증가를 하고 있다. 4차산업혁몽과 인공지능 시대에 중요한 것이 바로 데이터이다. 21세기는 데이터가 새로운 원유로서의 역할을 수행할 것으로 보인다. 이러한 데이터의 품질에 대한 관리가 매우 중요하다고 할 수 있다. 그러나 실무적인 차원에서의 연구는 진행이 되고 있으나 학문적 차원의 연구는 부족한 실정이다. 이에 본 연구에서는 전문가를 대상으로 데이터 품질관리에 영향을 미치는 요인에 대해 살펴보고 시사점을 제시하였다. 분석결과 데이터 품질관리의 중요도에는 차이가 있는 것으로 나타났다.

주제어 : 데이터, 데이터 품질, 데이터 품질관리, 데이터 평가 모델, 시스템 구축

Abstract This study is about the data quality management evaluation model. As the information and communication technology is advanced and the importance of storage and management begins to increase, the guam feeling for data is increasing. In particular, interest in the fourth industrial revolution and artificial intelligence has been increasing recently. Data is important in the fourth industrial revolution and the era of artificial intelligence. In the 21st century, data will likely play a role as a new crude oil. It can be said that the management of the quality of this data is very important. However, research is being conducted at a practical level, but research at an academic level is insufficient. Therefore, this study examined factors affecting data quality management for experts and suggested implications. As a result of the analysis, there was a difference in the importance of data quality management.

Key Words : Data, data quality, Data quality management, Data evaluation model, System construction

1. 서론

4차산업혁명시대에 중요한 키워드는 인공지능, 빅데이터, 클라우드, 네트워크, 사물인터넷, 모바일 등이 있다[1-3]. 이중에서도 중요한 것이 바로 빅데이터이다[4, 5]. 글로벌 기업 외에도, 글로벌 선진국을 위시한 국가에서는 공공 데이터들을 개방을 하고 있는 추세이다.

그러나 아직까지 데이터 관리 체계가 제대로 갖추어지지 않은 기업들과 기관들이 많이 존재를 하고 있다. 잘못 관리되고 잘못 입력된 데이터들로 인해 많은 사회적은

*Corresponding Author : Hyung-Sub Kim(kcy333@naver.com)
Received April 27, 2020 Revised May 29, 2020
Accepted July 20, 2020 Published July 28, 2020

비용과 문제를 초래하고 있다.

가트너에 의하면[6], 기업들이 매년 불량의 데이터를 처리하고 소모하는데 드는 비용이 평균적으로 1500만달러에 이른다고 하였다. 즉, 기업들의 약 60%는 불량 데이터를 측정 조차도 하지 않는다고 한다. 즉, 데이터 관리에 대해서는 중요하게 생각을 하지 않기 때문에, 데이터 품질 곤리 또한 관심을 기울이지 않고 있는 실정이다.

예를들면, 보건복지부에서는 2010년 사회복지통합망 구축하였으나, 사망자 정보를 잘못 관리하여 사망자 32만여명에게 639억원의 복지 급여를 잘못 지급을 하였다. 카드사에서는 연말정산시 신용카드에 대한 사용금액 누락되었다. 이로인해 290만명이 1,631억원의 피해를 입었다. 혼다가 DB 오류로 인해 미국에서만 766억원 벌금을 납부하였다[7]. 위에서 보는 바와 같이 데이터 품질 관리를 잘못하면 막대한 비용을 지불할 수가 있다. 또한 간접적으로는 사회에 많은 악영향을 미치고 있다.

4차산업에 필수적인 데이터의 융·복합, 민관 간 데이터의 자유로운 유통, 빅데이터 활용 활성화 등을 뒷받침할 수 있도록 공공데이터 표준화 및 품질 고도화가 필요할 것으로 보이며, 이를 통해 국가의 경제성장과 경쟁력 향상에 많은 도움이 될 것으로 보인다.

지금까지 데이터 품질관리가 시스템, 정보, 서비스 품질에 유의한 영향을 준다는 연구는 많이 수행이 되었다. 또한 데이터 품질관리시스템 설계 및 체계, 데이터 아키텍처, 관련 사례에 대한 연구들이 주를 이루고 있는 실정이다.

또한 대분의 연구에서는 이러한 점을 간과하고 있으며, 학문적 차원에서 보다는 실무(현장)에서 연구가 진행이 되고 있다. 물론 일부 연구에서는 데이터 품질이 기업의 경영성과에 미치는 영향을 파악하거나 직접 시스템을 구착하고 이를 검증하는 등의 연구가 일부 진행이 되고 있다. 그러나 대부분 글로벌 기업(IBM, MS, 구글, 오라클) 등에서 데이터에 대해 관심을 가지고 이를 관리할 수 있는 다양한 방법에 대해서 연구를 진행하고 있는 추세이다.

이에 본 연구에서는 데이터 품질관리의 중요요인을 도출하고자 한다. 이를 위해 데이터베이스 품질관리 지침에서 소개가 된 품질관리 요소 중에서 상위 중요요소 5개를 도출하고 전문가를 대상으로 설문조사를 실시하고 이를 통해 중요도를 산출하였다.

데이터의 품질관리 평가 요인에 대한 도출은 품질관리의 평가뿐만 아니라, 품질관리의 행위에 대한 가이드라인을 제시 해 줄 것이다.

2. 이론적배경

2.1 데이터 품질

데이터 품질에 관하여 다양한 연구가 진행되었으며, 연구자마다 조금씩 다른 관점의 정의를 내리고 있는데 선행연구들을 종합해보면 데이터 품질에 대한 가장 고전적인 정의는 '사용자의 목적에 적합하게 사용 가능한 수준'으로 볼 수 있다. 사용 적합성은 적절한 데이터 품질 수준은 상황에 따라 달라진다는 것을 의미한다.

Miller(1996)는 데이터 품질은 사용자에 의해 데이터가 어떻게 인식되고 사용되는가에 달려있다고 하였다[8]. Redman(2001)은 데이터는 운영, 의사결정, 계획을 하는데 있어 의도한 대로 사용하는데 적합하도록 결함이 없는 우수한 품질이어야 한다고 정의하였다[9]. 한국데이터베이스진흥원의 데이터 품질 가이드라인에서는 "데이터 품질이란 조직의 목적 달성을 위해 관리되는 데이터가 조직 구성원, 고객 등 데이터 이용자의 만족을 충족시킬 수 있는 수준을 의미한다"고 정의하였다[10].

2.2 데이터품질관리

최근들어 데이터의 품질관리에 대한 중요성이 증가를 하고 있다. 효율적인 경영을 위한 핵심 요소로서 데이터 품질관리가 떠오르고 있다. 데이터 품질관리에 관련된 선행연구들을 살펴보면, 데이터 품질관리를 협의의 개념과 공의의 개념으로 정의를 하고 있다.

먼저 조직에서 보유한 DB에 저장되어 있는 데이터를 수집, 처리, 보관, 분석하는 동안 무결성을 보장하는 비즈니스 프로세스로 정의하는 협의적 차원의 개념이 있다. 또한 데이터 관리 비전, 목표, 전략, 데이터 관리 원칙과 기준, 데이터 관리 절차 등을 모두 포괄하는 데이터 관리(거버넌스) 체계로 정의하는 광의적 차원의 개념이 있다[11, 12].

Table 1. Success factors for data quality management

Researcher	Success factors for data quality management
Firth(1996) [13]	Data quality management area identification and policy composition, data quality management purpose setting, managers and working group support
Segev & Zhao(1996) [14]	Recognize the importance of data quality at the organizational level, create information flows and process maps, identify data quality problems and locate them in the process map, and identify technologies and practical cases that can solve data quality problems

Wang et. al.(1998) [15]	Management of information from a product (product) point of view, understanding the information demand of customers
English(1999) [16]	Accurately understand what data quality improvement is and why it should be done, effectively improve data quality, improve data quality on the right issues, reward systems, training and communication, managers' interest and support for data quality improvement
Xu et. al.(2002) [17]	Training, management support, organizational structure, change management, employee relations
Otto et al.(2007) [18]	It is important to have a data governance foundation that defines the clear roles and responsibilities of data quality management.

3. 연구설계 및 조사 방법

데이터 품질 관리의 중요 요인을 도출하고 도출된 요인의 우선순위를 분석하기 위해 사용이 가능한 다양한 분석비법에 대해 살펴보고자 한다. 본 연구에서는 김영기 외(2010)와 박성택외(2010)의 연구에서 고려한 우선순위 분석 방법에 대해 살펴보고자 한다.

3.1 우선순위 결정 방법

우선순위 결정 방법에는 평점법, 다속성 효용이론, 델파이, AHP기법이 있다. 평점법은 우선순위를 결정하는 일반적인 방법중의 하나이다. 체크리스트 법을 논리적으로 확장시킨 모형이라고 할 수 있다. 그러나 평가과정에서 주관적일 가능성이 높기 때문에 신뢰성이 낮다는 단점이 있다[19].

다속성 효용이론은 복잡한 의사결정 과정에 대한 우선순위를 결정하는데 유용하게 사용되는 방법론이다. 그러나 평가기준의 수가 증가하면 증가할수록 우선순위 도출 작업이 복잡해 진다는 단점이 있다[20].

델파이는 미래에 대한 예측 기법 중의 하나로 많이 활용이 되고 있는 방법 중의 하나이다. 그러나 비과학적 이론이라는 비판을 받을 수 있다는 단점이 있다. 그러나 현재 시점에서 의사결정을 돕는다는 점에서 의의를 가지고 잇는 방법 중의 하나이다. 또한 전문가를 대상으로 정량이 아닌 정성적인 평가 방법으로서의 가치가 충분하다는 장점이 있다[21]. AHP 기법은 다수의 대안들의 우선 순위를 결정하는 방법론이다. 즉, 다 기준 의사결정의 방법으로 여러 대안 중에서 최적의 대안을 선택하는 방법으로 우선순위 도출방법에서 가장 많이 활용이 되고 있는 방법이다. 이에 본 연구에서는 AHP분석 방법을 사용하였다.

3.2 자료수집 및 분석 방법

AHP 조사는 2020년 3~4월에 걸쳐 실시하였으며, 설문대상은 각 기관의 담당자들에게 설문조사를 실시하였다.

응답대상자들은 현재 기관의 실무자들과 데이터 품질관리를 개발하는 개발자들로 구성이 되어 잇다. 총 응답자 13명 가운데, 불성실한 응답을 제외한 9명을 최종분석에 사용하였다. 쌍대비교행렬 A의 성분 a_{ij} 값들이 일관성을 크게 벗어나지 않는 한 $\lambda_{\max}$ 가 n 에 가까운 값을 갖는 성질을 이용하여 $A \cdot W = \lambda \cdot W$ 를 이용하여 가중치 W 를 추정할 수 있다. 이와 같은 방식으로 상위 평가기준(기술성, 사업성) 뿐만 아니라 각 하위 평가기준의 가중치, 복합가중치 등을 산출 할 수 있다[22].

4. 분석결과

4.1 분석결과

분석결과는 다음과 같다. 먼저 정확성이 1위로 나타났다. 데이터 품질관리 지침에 따르면, 정확성은 정확한 데이터 제공을 위해서는 데이터들이 입력 단계부터 오류가 없이 입력이 되어야 한다는 것이다. 또한 저장된 데이터가 사전에 정의된 기준에 맞도록 유효한 값의 범위와 형식을 갖추어야 하고, 현재 데이터베이스에 저장된 데이터가 현실에 가장 가까운 최신의 값을 반영하고 있어야 한다는 것을 의미한다.

빅데이터의 5V에서 중요한 것이 바로 정확성이다. 앞서 살펴본 바와 같이 정확하고 신뢰서 있는 데이터가 매우 중요하기 때문에 이러한 결과가 나온 것으로 보인다.

Table 2. Deducing the importance of data quality management factors

Factor	Weight	Ranking
Readiness	0.159	5
completeness	0.168	4
consistency	0.207	2
accuracy	0.294	1
Security	0.172	3

2위는 일관성으로 나타났다. 일관성은 같은 의미를 가지는 데이터들을 논리적인 속성의 단위, 물리적인 컬럼의

단위에서 일관된 이름과 형식을 가지도록 표준을 준수한다. 공공 데이터의 활용을 위해서는 공유 및 연계되는 데이터들은 하나의 누락도 없이 상호간에 일관성을 유지하여야 한다는 것이다. 데이터들이 일관성이 없으면 올바른 활용을 하지 못하기 때문에 2위로 그 중요도가 나타났다.

보안성은 3위로 나타났다. 보안성은 지속적으로 데이터들의 품질 확보를 위해 운영이 되고 있는 데이터들의 관리의 주체가 관리되어야 한다. 즉, 권한에 따라 데이터들에 대한 접근이 통제가 되어야 한다. 또한 개인정보와 같은 중요한 데이터들은 암호화 등에 대한 보안적인 조치가 이루어져야 한다.

완전성은 4위로 나타났다. 완전성은 데이터들의 저장소인 데이터베이스(DB)를 구축하는데 있어서 논리적 설계와 물리적 구조를 가져야 한다. 또한 업무의 요건에 맞도록 공공 데이터들이 저장되도록 해야 한다.

준비성은 5위로 나타났다. 준비성은 데이터들의 품질에 대한 관리를 위해서 기본적으로 관리를 해야 하는 정책, 규정과 조직 및 절차를 마련하고, 최신 내용 등으로 데이터들이 충실하게 관리되어야 한다.

4.3 시스템 구축을 통한 테스트

본 연구에서는 도출된 결과를 토대로 하여 시스템을 구축하고 이를 통한 실증 테스트를 진행하였다.

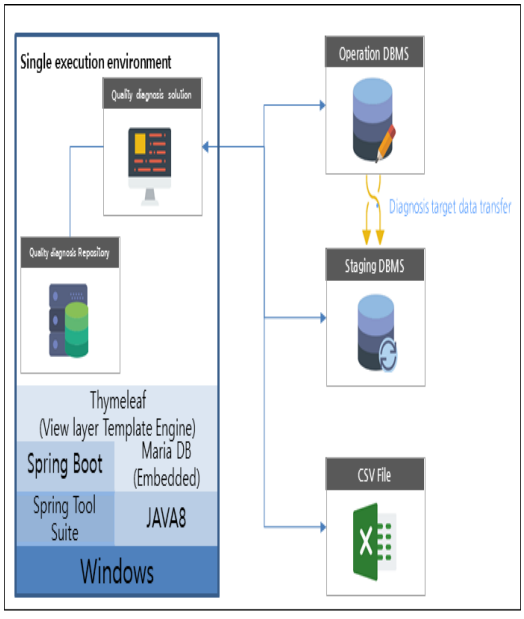

Fig. 1. Data quality management system

품질진단 대상인 DBMS는 운영 DBMS 또는 진단 대상 데이터를 별도로 저장해 놓은 Staging DBMS를 대상으로 하였다.

CSV 파일의 데이터를 진단할 수 있으며, 품질 진단 솔루션은 별도의 서버 설치 없이 단독으로 수행 가능 하도록 구성하였다.

동작 환경의 보안 구성에 영향을 받지 않도록 구성하고, 품질 진단 솔루션의 Repository DBMS(Maria DB)를 별도로 구성하였다.

컬럼구조 분석은 그림에서 보는 바와 같다.

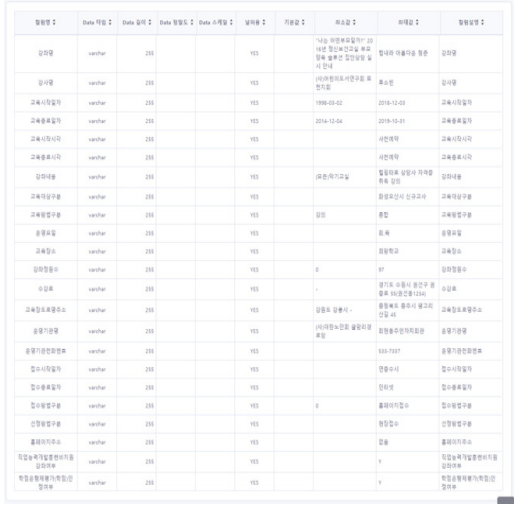

Fig. 2. Column structure analysis

진단 대상으로 설정한 DBMS 데이터 구조(테이블 구조)를 확인 하는 화면이다. 설정된 DBMS의 테이블 목록과 테이블 별 컬럼 정보를 확인(컬럼 명 / 데이터 타입, 길이 / 최대,최소 값 등)이 가능하다. 또한 접속한 계정의 권한에 따라 조회 가능한 스키마(데이터베이스) 목록과 테이블 목록을 조회 가능하다.

프로파일링 분석은 그림에서 보는 바와 같다. 프로파일링 실행 결과를 확인이 가능하다. 또한 기본적인 DBMS 정보와 테이블 명, 작업시간을 분석 개요에서 확인할 수 있다. 분석 결과는 컬럼 별로 분석 항목으로 제공하는 특징을 가지고 있다. 기본분석은 Data, Unique, Duplicate, Distinct, NULL, Blank Count를 확인이 가능하다. Value Frequency는 데이터 빈도 상위 10개 기준으로 조회한다.

기본 패턴 분석은 컬럼 항목별 선택한 분석방식에 따라 %로 매칭 정보를 확인한다. 불일치 정보와 일치 정보

는 불일치,일치 정보 다운로드 버튼을 클릭하여 CSV 파일로 다운 가능이 가능하다.

기존 시스템과 차별성은 메타 데이타 검증 및 대용량 데이터 베이스의 처리 능력이다. 또한 품질관리 된 데이터에 대해서 생명주기 관리가 가능하다.

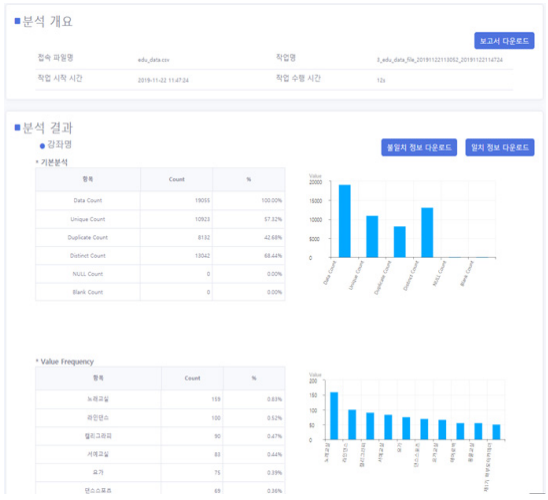

Fig. 3. Profiling analysis

프로파일링 실행 결과를 리포트로 출력 한다.

Fig. 4. Result report

진단대상정보로 데이터베이스의 스키마정보 및 테이블 컬럼 별 진단항목 건수, 항목 별 진단 컬럼 건수 를 제공한다. 종합현황에서 기본적인 DBMS 정보와 테이블명, 작업시간을 분석 개요에서 확인한다.

기본분석, Value Frequency, 기본패턴, 사용자 정의 부분으로 분석 항목 제공한다. 분석 항목별 분석결과를 차트 및 그래프로 제공이 가능하다.

5. 결론

본 연구는 데이터베이스의 품질 요인에 대한 연구로 데이터품질관리 매뉴얼에 기반을 하여 중요한 요인을 도출하여 분석을 수행하였다.

분석결과는 다음과 같다. 데이터 품질관리의 중요도를 도출한 결과, 정확성, 일관성, 보안성, 완전성, 준비성 순으로 그 중요도가 나타났다. 또한 분석결과를 토대로 시스템 구축을 통한 데이터 품질관리의 중요도를 실증 테스트하였다.

실무적 시사점으로는 본 연구에서 사용한 요인의 중요도를 통해, 시스템을 개발하는 개발자들에게는 본 연구결과를 활용할 수 있는 가이드라인을 제공할 것으로 보이며, 기관의 담당자들은 데이터베이스 품질관리의 중요성을 인식하고 이를 통해 데이터베이스 품질관리 지침에 따른 체계를 구축하는데 도움을 줄 수 잇을 것으로 보인다.

학문적 시사점으로는 본 연구를 통해 품질관리의 핵심 요인에 대해 중요성을 도출 할 수 있었다. 요인의 도출을 통해 세부 요인의 추가적 도출과 요인 과 요인간이 관계성 및 영향을 미치는 사항에 대해 추가적인 연구 가 필요해 보인다. 다만, 데이터 품질관리 영역이 전문화된 영역이어서 전문가 집단을 통한 조사에 어려움이 있다.

향후 연구에서는 설문조사 대상자를 확대할 필요가 있을 것으로 보이며, 데이터베이스 품질관리 시스템을 개발하고자할 때, 고려해야 할 사항을 사전에 정의하고, 기존의 데이터베이스 품질관리 시스템에서 갖추지 못하고 간과하였던 요인들을 도출해 낸다면 의미가 있을 것으로 보인다.

REFERENCES

[1] Park, S. T., Kim, D. Y. & Li, G. (2020). An analysis of environmental big data through the establishment of

emotional classification system model based on machine learning: *focus on multimedia contents for portal applications. Multimedia Tools and Applications,* 1-19.

[2] Park, S. T., Li, G. & Hong, J. C. (2018). A study on smart factory-based ambient intelligence context-aware intrusion detection system using machine learning. *Journal of Ambient Intelligence and Humanized Computing,* 1-8.

[3] Park, S. T. & Oh, M. R. (2019). An empirical study on the influential factors affecting continuous usage of mobile cloud service. *Cluster Computing, 22(1),* 1873-1887.

[4] Park, S. T., Jung, J. R. & Liu, C. (2019). A study on policy measure for knowledge-based management in ICT companies: *focused on appropriability mechanisms. Information Technology and Management,* 1-13.

[5] Park, S. T., Lee, S. W. & Kang, T. G. (2018). A study on the trend of cloud service and security through text mining technique. I*nternational Journal of Engineering & Technology, 7(2.33),* 127-132.

[6] Gartner, https://www.gartner.com/en

[7] Korea Institute for Health and Social Affairs. (2013). *The basic direction of social security in the next 5 years and Finding core tasks.*

[8] Miller, J. S. (1996). U.S. Patent No. 5,506,984. Washington, DC: *U.S. Patent and Trademark Office.*

[9] Redman, T. C. (2001). Data quality: the field guide. *Digital press.*

[10] Korea Data Agency. (2012). *Data Quality Management Guidelines.*

[11] An, H. J. (2016). A Business Performance Study of Data Quality Management for Big Data Adoption - *Focused on Corporate Data Quality Management Process* -, Kookmin University.

[12] Park, G. H. (2017). *The Determinant for the Usage of Big Data in Administrative Service : mainly on the Quality Control of Data,* Keimyung University.

[13] Firth, C. P. (1996, October). Data Quality in Practice: *Experience from the Front Line. In IQ* (pp. 65-71).

[14] Segev, A. & Zhao, J. L. (1996). Rule activation techniques in active database systems. *Journal of Intelligent Information Systems, 7(2),* 173-194.

[15] Wang, H., Long, Q., Marty, S. D., Sassa, S. & Lin, S. (1998). *A zebrafish model for hepatoerythropoietic porphyria. Nature genetics, 20(3),* 239-243.

[16] English, L. P. (1999). *Improving data warehouse and business information quality.* methods for reducing costs and increasing profits (Vol. 1). New York: Wiley.

[17] Xu, Y., Olman, V. & Xu, D. (2002). Clustering gene expression data using a graph-theoretic approach: *an application of minimum spanning trees. Bioinformatics, 18(4),* 536-545.

[18] Otto, B., Wende, K., Schmidt, A. & Osl, P. (2007). *Towards a framework for corporate data quality management.*

[19] Kim, Y. K., Lee, S. J. & Park, S. T. (2010). Selection of important factors for Patent Valuation using Delphi Method. *Entrue Journal of Information Technology, 9(1),* 7-17.

[20] Park, S. T., Lee, S. J. & Kim, Y. K. (2011). Weight Differences of Patent Valuation Factors by Industries. *Journal of Digital Convergence, 9(3),* 105-116.

[21] Kim, Y. K., Lee, S. J. & Park, S. T. (2011). Establishing the Importance Weight of Patent Valuation Criteria for Product Categories through AHP Analysis. *Entrue Journal of Information Technology, 10(1),* 115-127.

[22] Lee, S. J., Kim, Y. K. & Park, S. T. (2013). Appropriability Mechanism Strategy for Domestic IT Manufacturing Companies. *Journal of Digital Convergence, 11(11),* 233-242.

김 형 섭(Hyung-Sub Kim) **[정회원]**

- 2003년 2월 : 한양대학교 화학과 경영학과(이학사, 경영학사)
- 2005년 2월 : 한양대학교 경영학과 (경영학석사)
- 2013년 9월 : 한양대학교 경영학과 (박사 수료)
- 2015년 3월 ~ 2017년 8월 : 선문대학교 경영학과 겸임교수
- 2005년 3월 ~ 현재 : 상상스토리(주) 대표이사
- 관심분야 : 경영데이터분석, 경영시뮬레이션게임
- E-Mail : kcy333@naver.com