한국빅데이터학회지
제3권 제2호, 2018, pp. 11-18

머신러닝을 이용한 빅데이터 도메인 자동 판별에 관한 연구*

A Study of Big Data Domain Automatic Classification Using Machine Learning

공성원† · 황덕열

㈜위세아이텍

요 약

본 연구는 빅데이터 품질 진단의 핵심 요소인 도메인 기반 품질 진단을 위한 도메인 자동 판별에 관한 연구다. 빅데이터의 가치와 활용도의 증가와 4차 산업혁명의 대두로, 법률, 의료, 금융 등 IT와 융합된 다양한 분야에서 빅데이터를 활용하여 새로운 가치를 창출하려는 노력을 진행중이다. 하지만, 신뢰도가 낮은 데이터에 기반한 분석은 과정과 결과 모두에서 치명적인 문제를 발생하며, 분석 결과에 따른 판단 또한 신뢰하기 어려워 진다. 이처럼 신뢰도가 높은 데이터의 필요성 또한 증가하였지만, 데이터의 품질 확보에 대한 연구와 그에 대한 결과는 미비하다. 본 연구는 데이터 품질 향상을 위한 진단 평가의 핵심적 요소인 도메인 기반 품질 진단에서, 수작업으로 진행되었던 도메인 판별 작업을 머신러닝을 이용하여 자동화 함으로써, 작업시간을 단축하는 것을 목표로 한다. 데이터 베이스에 저장된, 도메인이 판별되어 있는 데이터의 특성에 관한 정보들을 추출하여 변수화하고, 이를 머신러닝을 이용하여 도메인 판별을 자동화 한다. 이를 빅데이터 품질 진단에 활용하고, 품질 향상에 기여하도록 한다.

■ 중심어 : 빅데이터, 데이터 품질 진단, 도메인, 머신러닝, 랜덤 포레스트

Abstract

This study is a study on domain automatic classification for domain - based quality diagnosis which is a key element of big data quality diagnosis. With the increase of the value and utilization of Big Data and the rise of the Fourth Industrial Revolution, the world is making efforts to create new value by utilizing big data in various fields converged with IT such as law, medical, and finance. However, analysis based on low-reliability data results in critical problems in both the process and the result, and it is also difficult to believe that judgments based on the analysis results. Although the need of highly reliable data has also increased, research on the quality of data and its results have been insufficient. The purpose of this study is to shorten the work time to automizing the domain classification work which was performed from manually to using machine learning in the domain - based quality diagnosis, which is a key element of diagnostic evaluation for improving data quality. Extracts information about the characteristics of the data that is stored in the database and identifies the domain, and then featurize it, and automizes the domain classification using machine learning. We will use it for big data quality diagnosis and contribute to quality improvement.

■ Keyword : Big Data, Data Quality Diagnosis, Domain, Machine Learning, Random Forest

2018년 11월 06일 접수; 2018년 11월 08일 수정본 접수; 2018년 12월 31일 게재 확정.

* 이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (2017-0-00163, 빅데이터 품질평가 도구 개발)

† 교신저자 swkong@wise.co.kr

Ⅰ. 서 론

현대 사회에서 빅데이터의 가치는 높아져서 IT 뿐 아니라, 공공기관, 법률, 의료, 금융 등 다양한 분야에서 빅데이터를 활용하여 새로운 가치를 창출하고 있다. 정부에서도 이러한 움직임에 발맞춰, 빅데이터 관련 산업에 많은 투자를 시작하였다. 특히 2017년 이후, 4차 산업혁명의 활성화를 위하여 공공데이터를 개방하였다.[9]

빅데이터를 이용하여 새로운 가치를 창출하기 위해서는 신뢰도가 높은 데이터가 전제되어야 한다. 낮은 신뢰도 데이터 기반의 분석은 분석 과정에서부터 문제가 생길 뿐더러, 분석 결과에 따른 판단에도 오류를 범할 수 있다. 때문에 전 세계적으로 민간부분의 데이터 신뢰성과 품질확보를 위해 년간 6000억 달러 이상의 비용을 소비하고 있으며, 품질관리 수준을 평가하기 위한 지표 등에 관한 연구들도 진행 중이다.[8] 우리나라의 경우 공공기관을 중심으로 공공데이터의 품질을 높이기 위하여 투자를 시행하고 있다.[5][10]

데이터 품질에 관한 이슈는 데이터 마이닝으로 인한 가치 창출과 인공지능 산업 전반에 걸친 문제가 될 수 있다. 이번 연구에서는 데이터 품질 진단 방법 중, 도메인 기반 데이터 품질진단에서, 기존에 제안하였던 머신러닝 기반 도메인 자동 분류 시스템의 문제점을 개선하고 발전시킨 방법의 연구를 진행하였다.

Ⅱ. 빅데이터와 머신러닝

빅데이터는 크기가 크고(Volume), 빠른 처리 속도(Velocity)와 높은 다양성(Variety)의 특징을 가지고 있는 데이터라고 정의하고 있다.

저장 기술의 발전과 비용의 하락, 그리고 인터넷과 모바일 시대에 들어섬에 따라 기술적으로 빅데이터의 탄생과 유지가 가능해 졌다. 이를 많은 산업계에서 이용하기 시작하여, 사용자 및 소비자 행태 정보를 적극 수집, 분석하여 경영전략에 사용하기 시작하였다. 또한 학문 분야에서도 이를 사용하기 위한 방법과 도구 기술 등을 지속적으로 연구 개발하면서 빅데이터의 발전을 돕고 있다.[11]

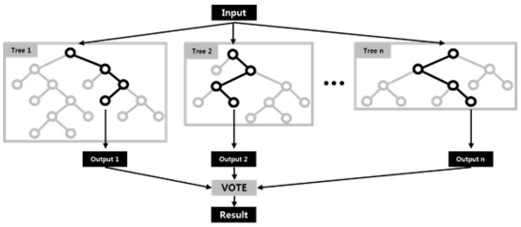
빅데이터의 성장은 데이터를 분석하기 위한 기법과 기술에 관한 연구들을 활성화 시켰다. 빅데이터의 성장과 더불어 특히 컴퓨팅 파워의 지속적인 발전은 한때는 학문분야에 그쳤던 머신러닝이라는 분야를 상업적, 실용적으로 이용하려는 노력을 지속화 시켰다.

머신러닝 (Machine Learning)이란 컴퓨터 과학의 한 분야로 컴퓨터를 인간의 학습 능력과 같은 기능으로 실현하고자 하는 기술이다. 경험적 데이터셋을 특정한 모델로 적합화하여, 학습시키고 예측을 수행하면서 스스로의 성능을 향상시킬 수 있다. 학습된 머신러닝 알고리즘들의 모델은 엄격하게 정의된 조건에 의한 명령을 수행하는 것이라기 보다, 입력 데이터들을 기반으로 예측이나 결정이 가능하도록 구축되었다[12].

Ⅲ. 랜덤 포레스트

랜덤 포레스트 (Random Forest)는 머신러닝 알고리즘의 하나로 의사결정 트리를 기반으로 한 앙상블 학습방법이다. 의사결정 트리는 객체를 분류하거나 예측하는 방법으로, 트리의 각 노드에서의 조건들을 이용하여 구분값들을 분류하는 알고리즘이다. 이러한 의사결정 트리들을 여러 개로 합쳐서 만든 알고리즘이 랜덤 포레스트 기법이다.

<그림 1>의 상자안의 노드들 같이 의사결정 트리를 여러 개를 만들어 학습시키고, 각각의 학습된 의사결정 트리에서 나온 예측 결과들을 투표하여(Voting), 가장 많은 숫자가 나온 결과



〈그림 1〉 랜덤 포레스트 알고리즘

를 랜덤 포레스트의 결과로 도출해내는 알고리즘이다[3]

Ⅳ. 데이터 품질

데이터 품질이란 <데이터의 최신성, 정확성, 상호연계성 등을 확보하고, 이를 이용하여 사용자에게 유용한 가치를 줄 수 있는 수준>으로 정의 하고 있다. 데이터 품질 진단을 실시하는 목적은 데이터 품질을 체계적, 지속적으로 유지하고 향상시키기 위함이다.

특히 공공기관의 데이터 개방과 맞물려 데이터의 품질을 진단하고 향상하려는 노력에 많은 자원이 투자되고 있다. 데이터 품질을 평가하고 인증하기 위한 수단으로 DQC-V, DQC-M 등과 같은 인증 수단을 실시하고 있는데, 데이터 값에 대한 인증영역으로 도메인과 업무 규칙을 기준으로 데이터 자체에 대한 품질 영향 요소를 심사하고 인증하는 제도이다.[5]

업무 규칙은 업무와 관련된 모든 데이터의 규칙을 말한다. 업무 규칙은 조직에서 데이터 품질을 지속적으로 관리하기 위해 사용하는 데이터 측정 규칙이며 데이터 값이 정확하기 위한 조건 표현이다.[6] 업무 규칙 기반 품질 진단을 하기 위해서는 업무 규칙을 파악하고, 조건 또는 제약을 설정하여 규칙을 SQL 등을 이용하여 실제 운영 데이터베이스에 적용한다. 적용한 데이터베이스에서 오류데이터를 추출하고 오류율을 확인함으로써 품질을 진단한다.

도메인 기반 데이터 품질 진단은 데이터에 대한 프로파일링 기법을 사용하기 위한 작업으로, 도메인을 분류된 컬럼을 이용하여 각각의 도메

〈표 1〉 도메인 분류

도메인 분류	도메인 예시	점검내용
번호	주민등록번호, 사업자등록번호, 우편번호, 고객번호, 계좌번호	번호 관련 데이터의 패턴 및 체크비트 진단
금액	금액, 세금, 가격, 단가, 비용, 요금, 잔액, 총액	금액 관련 데이터의 허용범위 진단
명칭	명, 주소, ID, 장소, 고객명, 영문 고객명, URL, 이메일, IP	명칭 관련 데이터의 패턴 및 길이 진단
수량	건수, 매수, 회차, 개수, 거리, 규모, 길이, 무게, 속도, 횟수, 평형, 면적, 온	수량 관련 데이터의 허용범위 진단
분류	여부, 유무, 구분, 상태	분류 관련 데이터의 표준정의 값 진단
날짜	년월, 년, 년월일, 시, 분, 초, 일, 반기, 분기	날짜 관련 데이터의 허용범위 및 유효값 진단
율	금리, 이율, 비율, 환율, 백분율	비율(%) 관련 데이터의 허용범위 진단
내용	내용, 비고, 설명, 정보, 요약	내용 관련 데이터의 적용언어 패턴 진단
코드	개별코드, 통합코드	코드 관련 데이터의 코드값 진단
키	일차키, 외래키	키 관련 데이터의 참조 무결성 진단
공통	데이터 표준화	데이터 표준 준수여부 진단

인의 특성에 맞게 컬럼 분석, 날짜 분석, 패턴 분석 등을 수행하여 평가한다. 이를 수행하기 위해서는 해당 칼럼에 대한 도메인이 분류가 되어 있어야한다. 도메인은 <표 1>과 같다.

도메인 분류 작업의 경우, 사용자가 데이터를 일일이 확인하고 수작업으로 진행하였기 때문에, 휴먼에러가 자주 발생하였고, 작업시간 역시 많은 시간이 소요되는 문제점을 가지고 있었다.

Ⅴ. 연구 및 결과

5.1 연구목적

앞서 설명 하였듯, 도메인 기반 데이터 품질을 진단하기 위해서는 많은 물리적, 시간적, 인적 재원을 투자해야 했다. 때문에 ㈜위세아이텍에서는 이러한 자원의 소모를 줄이기 위한 연구를 진행하고 있으며, 그 일환으로 도메인 자동화 분류 시스템의 문제점을 제안하였다. 프로파일 기반 품질 진단 데이터를 이용하여, 의사결정 트리 알고리즘을 학습시켜 도메인 자동 판별 시스템을 구성하였다.[1]

<표 2>는 기존의 도메인 자동화 분류 시스템에서사용하였던 변수의 목록이다.[1] 프로파일 품질 진단의 결과값으로 도출되는 변수들과 칼럼의 논리명, 물리명등에서 도출한 파생 변수들이다. 하지만 파생변수를 도출해 내기 위해서는 데이터베이스 또는 해당 테이블의 데이터 표준화가 전제 되어야 있어야 하는 문제점이 있었다. 예를 들어, 표준화가 되어있지 않은 데이터베이스, 테이블의 경우, 물리 칼럼명이 형식을 가지고 있지 않고, 논리 칼럼명은 없는 경우가 대부분이다. 때문에 실제로 표준화를 시키지 않고는 해당 파생 변수가 널 값을 가지고 있는 경우가 대부분이다. 그 데이터를 이용하여 데이터베이스의 칼럼 도메인을 자동 판별하였을 경우, 정확도가 감소하는 문제점을 가지고 있었다. 또한, 학습 모델로 제안하였던 의사결정 트리 알고리즘의 과적합 문제점도 내재되어 있었다.

5.2 변수 설정 및 수집

표준화되어 있지 않은 데이터베이스에서 도메인 판별을 수행하기 위하여, 기존의 변수를 수정하고 추가하였다. 그 과정에서 반드시 표준화가 되어 있어야 추출이 가능한 파생 변수들은 삭제하고, 데이터의 대표값을 파생 변수화 하였

〈표 2〉 기존 파생 변수 정의[1]

변수	설명
데이터 타입	INT, CHAR, VARCHAR 등 같은 데이터 값을 구분할 수 있는 변수
논리 칼럼명 접미사	고객명, 상품명, 주민등록번호 매출액과 같이 사림이 인식할 수 있는 칼럼의 정하는 명칭이 논리 칼럼명이고 접미사는 논리 칼럼명의 마지막 형태소를 의미
물리 칼럼명 접미사	CUST_NAME, PRODUCT_ID와 같이 컴퓨터가 이해할 수 있는 칼럼의 이름을 의미하며 NAME, ID와 같이 마지막에 사용된 단어 또는 '_' 와 같은 특수 기호로 구분되는 마지막 단어를 의미
소수점 포함 여부	숫자형 데이터 타입인 칼럼 중 데이터에 소수점 포함 여부에 따라 분류
날짜 데이터 여부	DATE, TIMESTAMP와 같이 날짜 데이터 타입을 포함하고, 문자형 데이터 타입 중에서도 날짜형으로 데이터가 들어가 있는지 여부
숫자 여부	INT, FLOAT 등과 같은 숫자 데이터 타입 여부
데이터 중복 제외 건수	고유한 데이터 건수를 의미
텍스트 200자 초과여부	문자형 데이터 타입 스키마 상의 길이 200자 이상 여부
텍스트 자릿수 변동 여부	칼럼에 존재하는 데이터의 자릿수 변동 여부

다. 파생 변수들은 다음과 같다.

〈표 3〉 파생 변수 정의

변수	설명
데이터 타입	INT, CHAR, VARCHAR 등 같은 데이터 값을 구분할 수 있는 변수
데이터 최대길이	칼럼 내의 데이터 중 최대 길이를 가지고 있는 데이터의 길이
데이터 최소길이	칼럼 내의 데이터 중 최소 길이를 가지고 있는 데이터의 길이
데이터 길이 변화	칼럼 내의 데이터 길이의 가변 여부
소수점 아래 길이	칼럼 내의 데이터들의 소수점 아래의 길이
날짜 형식 여부	데이터 타입이 아닌 날짜 포맷 데이터 여부
연락처 형식 여부	@, - 등 연락처 및 주소에서 사용하는 패턴을 이용한 데이터 존재여부
공백 비율	전체 데이터에서 공백이 차지하는 여부
엔터 포함 여부	칼럼 내의 데이터에서 줄바꿈이 일어났는지 여부
영어 작성 여부	데이터들이 영어로만 작성되었는지 여부
숫자 작성 여부	데이터들이 숫자로만 작성되었는지 여부
백단위 이하 비율	칼럼 내의 데이터 중 100단위 이하는 000으로 표기 된 비율
그룹화 비율	칼럼 내의 데이터 중 그룹화가 가능한 비율
PK 여부	칼럼이 Primary Key로 설정되었는 지 여부

표준화 여부와 상관 없이, 데이터베이스와 SQL문을 사용하여 추출할 수 있는 정보만을 변수로 선택하여 파생 변수화 하였다. 각각의 변수들의 선택기준은 도메인 분류 시에 점검 내용의 기준이 되는 사항들을 파생변수화 하였다.

데이터의 수집은 오픈 되어있는 공공데이터들을 이용하여 수집하고 라벨링을 수행하였다. [9] 라벨링은 <표>에 나온 도메인 영역 중 Key와 공통을 제외한, 번호, 금액, 명칭, 수(량), 분류(플래그> 날짜, 율, 내용, 코드에 더하여 사업분야에서 많이 다루고 요구되는 연락처를 추가한 총 10개의 영역으로 라벨링 하였다.

5.3 분류 모델 개발

랜덤 포레스트 모델은 과적합 추정 모델을 결합해서 과적합의 효과를 줄일 수 있다. 랜덤포레스트 모델을 사용하여, 의사결정 트리에 잠재되어 있는 과적합 문제를 방지하였으며, 늘어난 변수 개수에 대응할 수 있다. [2]

10개의 도메인으로 구분한 데이터셋을 랜덤포레스트 모델을 사용하여 학습하였다. 학습된 모델에서 변수의 영향도를 확인한 결과는 다음과 같다.

〈표 4〉 파생변수 영향도

변수	영향도
데이터 길이 변화	19.60%
날짜 형식 여부	15.40%
데이터 최대길이	15.30%
그룹화 비율	12.70%
데이터 최소길이	8.90%
데이터 타입	7.70%
공백 비율	5.60%
연락처 형식 여부	4.40%
숫자 작성 여부	4.30%
백단위 이하 비율	3.20%
PK 여부	2.10%
엔터 포함 여부	0.40%
소수점 아래 길이	0.20%
영어 작성 여부	0.20%

영향도가 1%가 되지 않는, 소수점 아래 길이, 엔터 포함 여부, 영어 작성 여부를 제외하고 학습모델로 구성하여, 최종적으로 11개의 변수를 사용하여 자동화 시스템을 구성하였다.

〈표 5〉 변수

데이터타입	최대길이	최소길이	길이 가변	날짜 형식	연락처 형식	공백비율	숫자 작성	백단위 비율	그룹화 비율	PK 여부	도메인
VARCHAR2	1	1	N	N	N	0	N	0	0.69	N	플래그
NUMBER	3	1	Y	N	N	0	N	0	0.05	N	수
NUMBER	3	1	Y	N	N	0	N	0	48.62	N	번호
VARCHAR2	11	11	N	N	N	0	N	0	4.14	N	코드
VARCHAR2	20	9	Y	N	N	38.29	N	0	10.64	N	명칭
VARCHAR2	430	2	Y	N	N	60	N	0	66.67	N	내용
VARCHAR2	52	6	Y	N	N	48.33	N	0	35	N	명칭
BLOB	10000	500	Y	N	N	0	N	0	0	N	내용
DATE	8	8	N	Y	N	0	N	0	100	N	날짜
NUMERIC	18	1	Y	N	N	0	N	0	0	N	금액
VARCHAR2	9	2	N	N	Y	1.66	Y	0	48.33	N	연락처
VARCHAR2	29	2	N	N	N	15.38	N	0	46.15	N	명칭

Ⅵ. 연구 결과

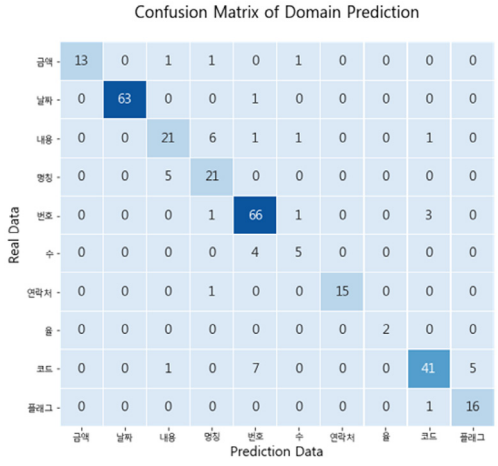
공공데이터로부터 수집된 3400개의 데이터셋을 라벨링하여 학습 및 예측을 실행하였다. <표 5>는 사용된 데이터셋의 Sample 데이터이다. 학습과 예측을 위한 데이터들은 8:2로 나누어 진행하였다.

〈표 6〉 알고리즘 결과

Accuracy	Precision	Recall
86.2%	86.8%	86.2%

정확도는 정밀도, 재현률 모두 85 %이상의 결과를 도출하였다. 수집된 데이터들의 숫자가 늘어나면 성능이 향상될 것으로 기대된다.

<그림 2>은 예측 데이터의 Confusion Matrix이다. 실제 데이터와 예측된 데이터들의 관계를 확인할 수 있다. 이 관계에서 <코드>와 <플래그> 관계들을 확인해보면, 두 도메인 사이에서만 구분을 못하는 경우 확인된다. <번호>와 <수>의 관계 또한 마찬가지이다. 데이터셋을 늘려 재학습을 시키지 않는 이상, 이 관계에서는 구분하기가 힘들다. 정확도를 높이기 위해서, 구분하기 힘든 <코드>와 <플래그>를 통합하여 나머지 도메인들과 같이 머신러닝을 이용하여 구분한 다음, 규칙을 이용하여 코드와 플래그를 구분하는 방법을 사용하면 정확도, 정밀도, 재현율을 높일 수 있을 것으로 기대 된다.



〈그림 2〉 Confusion Matrix

Ⅶ. 결 론

본 논문은 머신러닝을 사용하여 데이터 품질 진단 평가에 적용한 연구다. 도메인 품질 진단

평가를 위해서 수작업으로 진행되었던, 도메인 판별 작업을 머신러닝을 이용하여 자동화 함으로써, 작업시간을 단축하는 것을 목표로 하였다.

머신러닝을 사용한 도메인 자동 판별 알고리즘을 위해서 데이터베이스에서 얻을 수 있는 파생변수들을 제안하였고, 파생변수에 대해서 설명하였다. 그리고 학습 알고리즘에 대한 파생변수의 영향도를 평가하고, 축소하였다. 이를 이용하여 도메인 자동 판별 알고리즘을 구성하고 평가하였다.

추후, 연구과제의 목표로는 다양한 산업군에서 수집한 파생변수들을 수집하고, 학습 및 평가하여 도메인 통합, 머신러닝과 규칙을 통한한 알고리즘 개발 등에 대한 연구를 지속적으로 진행하여, 빅데이터 품질 평가를 위한 도메인 자동화 판별 도구를 개발하여 빅데이터의 품질향상에 기여할 예정이다.

참 고 문 헌

[1] 이진형, "머신러닝을 이용한 빅데이터 품질진단 자동화에 관한 연구", *한국빅데이터논문지*, 제2권 제2호, 2017

[2] Robert E. Schapire, "Random Forests", *Machine Learning*, 45, 5–32, 2001

[3] A Liaw, M Wiener, *Classification and regression by randomForest*, R news, 2002

[4] B.P.Weidema, M.S.Wesnæs, *Data quality management for life cycle inventories—an example of using data quality indicators*, Vol4, Issues 3–4, 1996, Pages 167-174

[5] 이상기, 채철주, 홍의경," 데이터 프로파일링과 정규 표현식 활용 비정형 과학기술 빅데이터 품질관리 방안", *한국콘텐츠학회논문지*, 제14권, 제12호, p486-793, 2014

[6] 명재호, 안희진 이창수, 김성현 임동진, 오경조, 이종규, 김선영, 최용준, 데이터 품질 가이드라인, 한국데이터진흥원, 2011

[7] 데이터 품질관리 지침, 한국데이터베이스진흥센터, 2006

[8] 데이터 산업 백서, 한국데이터진흥원, 2017

[9] 차경엽, 심광호, "공공부문 정보시스템 데이터의 신뢰성 점검기법 개발", *한국통계학회논문집*, 제17권, P745-753, 2010

[10] *데이터 분석 전문가 가이드*, 한국데이터베이스진흥원, 2016

[11] J. VanderPlas, *Python Data Science Handbook: Essential Tools for Working with Data*, 2016

[12] T.F. Cootes, M.C.Ionita, C.Lindner, P.Sauer, "Robust and Accurate Shape Model Fitting Using Random Forest Regression Voting", *Computer Vision – ECCV 2012*, pp 278-291, 2012

[13] 김선호, 이창수, "데이터 품질관리 프로세스 평가를 위한 프로세스 참조모델", *한국전자거래학회지*, 제18권, 2013

[14] Caballero, I., Caro, A., Calero, C., Piattini, M., "IQM3 : Information Quality, Management Maturity Model," Journal of Universal Computer Science Vol. 14, No. 22, pp. 3658-3685, 2008.

[15] ISO 8000-1 Data quality—Part1 : Overview, ISO, 2009

[16] Pipino, L. L., Lee, Y. W., Wang R. Y., "Data quality as-sessment", Communications of the ACM, Vol. 45, No. 4, pp. 211-218, 2002.

[17] Ryu, K. S., Park, J. S., Park, J. H., "A data quality management maturity model," ETRI Journal, Vol. 28, No. 2, 2006.

[18] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang, "Data Quality Assessment," Communications of the ACM, vol. 45, no. 4, Apr. 2002, pp. 211-218.

저 자 소 개

공 성 원(Kong Seongwon)

- 2013: 건국대학교 항공우주정보시스템공학과(학사)
- 2015: 건국대학교 항공우주정보시스템공학과(석사)
- 2016~현재 : 위세아이텍 DM사업부 선임
- 관심분야 : GNSS, 데이터마이닝, 이상값탐지, 인공지능

황 덕 열(Hwang Deokyoul)

- 1997년 : 성균관대학교 컴퓨터공학(석사)
- 2000년 ~ 현재 : 위세아이텍 DM사업부
- 관심분야 : 데이터관리, 빅데이터 품질, 인공지능