RESEARCH

KEPCO Journal on Electric Power and Energy
p-ISSN 2465-8111, e-ISSN 2466-0124
Volume 7, Number 1, June 2021, pp. 79-84
DOI 10.18770/KEPCO.2021.07.01.079

Automatic Cleaning Algorithm of Asset Data for Transmission Cable

지중 송전케이블 자산데이터의 자동 정제 알고리즘 개발연구

Jae-Sang Hwang, Sung-Duk Mun, Tae-Joon Kim, Kang-Sik Kim

Abstract

The fundamental element to be kept for big data analysis, artificial intelligence technologies and asset management system is a data quality, which could directly affect the entire system reliability. For this reason, the momentum of data cleaning works is recently increased and data cleaning methods have been investigating around the world. In the field of electric power, however, asset data cleaning methods have not been fully determined therefore, automatic cleaning algorithm of asset data for transmission cables has been studied in this paper. Cleaning algorithm is composed of missing data treatment and outlier data one. Rule-based and expert opinion based cleaning methods are converged and utilized for these dirty data.

Keywords: *Asset Data, Asset Management System, Data Cleaning, Legacy System, Missing Data Treatment, Outlier Data Treatment, Preprocessing, Transmission Cables*

I. Introduction

전세계적으로 전력설비의 노후화가 진행됨에 따라 많은 전력설비가 설계수명에 근접하거나 초과하여 운전되고 있고, 노후 전력설비의 관리를 위해 고장 위험성의 우선순위를 파악하여 고장 파급성이 높은 전력설비를 우선적으로 교체함으로써 경제적이고 효율적인 설비 투자가 이루어질 수 있도록 전력설비의 자산관리시스템을 도입 및 구축 중에 있다. 자산관리 기술은 ISO 55000 시리즈를 근간으로 하고 있으며, 여기서 자산관리를 위한 프로세스로 자산관리 체계, 자산관리 목표, 자산데이터 수집·분석, 자산관리 전략, 자산관리시스템 구축·운영, 자산관리시스템 평가를 Fig. 1과 같이 제시한다 [1]-[3]. 세계 전력회사별 자산관리시스템의 기술 격차는 자산데이터 품질과 RISK 평가 알고리즘의 정확성에 달려있다.

자산관리시스템은 운영하고 있는 Legacy System으로부터 데이터 연계를 통해 설비자산에 대한 고장확률과 고장영향을 종합평가하는 RISK 평가 알고리즘을 통해 설비 교체 우선순위를 정해 경영진의 투자 의사결정을 지원한다. 만약 Legacy System의 데이터가 부정확하다면 이러한 데이터를 이용해 RISK 평가 알고리즘에서 도출된 결과는 부정확할 수밖에 없다. 이에 따라 전력설비 교체 우선순위가 달라질 수 있고, 부정확한 투자계획이 수립될 수 있기 때문에 Legacy System의 데이터 품질이 매우 중요하고 이는 곧 자산관리시스템의 신뢰성으로 직결된다.

기업의 82%가 데이터는 가장 전략적인 자산이라는 점에 동의하고 있으므로 데이터의 고품질 관리는 필수요소라고 볼 수 있다 [4]. 전력설비의 설비정보, 점검정보, 고장정보 등은 다양한 Legacy System에서 데이터로 수집·저장되어 관리되고 있고, Legacy System간 연계가 되고 있는 실정이다. 지중송전케이블에 관련된 대표적인 Legacy System은 현재 송전운영시스템(TOMS), 송변전통합설비관리시스템(STOM), 변전소운전실적관리시스템(SOMAS)으로 구성되며, 이 중 TOMS과 STOM이 자산관리시스템에 연계되어 사용될 자산데이터를 주로 관리하고 있다. TOMS는 지리정보 기반의 신규설비 생성 및 이력관리를 통해 송전 설비관리 업무 전반을 관리하는 시스템으로써 설비 운영정보를 관리하는 시스템이며, STOM은 점검/정비 관련 송변전 분야별 분산된 시스템을 통폐합하여 순시/점검결과를 입력하여 관리한다.

자산관리시스템을 운영하기 위해서는 자산정보, 점검이력, 고장이력 등과 같은 데이터가 반드시 연계되어야 하기 때문에 데이터의 관리 및 운영시스템 여부가 자산관리 기술을 구축 및 운영할 수 있는지에 대한 지표가 되어 설비자산에 대한 데이터의 보유 여부가 매우 중요하다 [5]. Legacy System에서 생성된 모든 원시데이터(Raw data)는 자산관리시스템에 연계되므로 Legacy System의 데이터 품질이 자산관리시스템 내 RISK 평가 알고리즘 결과에 의한 설비자산 투자전략에 영향을 미칠 수 있어 고품질 데이터 확보가 요구되고 있다. 하지만, 자산정보나 현장에서의 점검 및 진단이력에 대한 수기 입력에 의해 오기입이 발생할 수 있어 데이터의 정제가 반드시 필요하다. 데이터의 정제를 하기 위해서는 생성된 데이터가 정확한 정보인지, 부정확한 정보인지를 판가름할 수 있는

Article Information

Manuscript Received January 12, 2021, Accepted January 19, 2021, Published online June 30, 2021

The Authors are with KEPCO Research Institute, Korea Electric Power Corporation, 105 Munji-ro Yuseong-gu, Daejeon 34056, Republic of Korea.

Correspondence Author: *Jae-Sang Hwang* (*jshwang@kepco.co.kr*)

ORCID: *0000-0002-7394-5318* (*K. S. Kim*)

Fig. 1. Structure of asset management system based on ISO 55000 series.

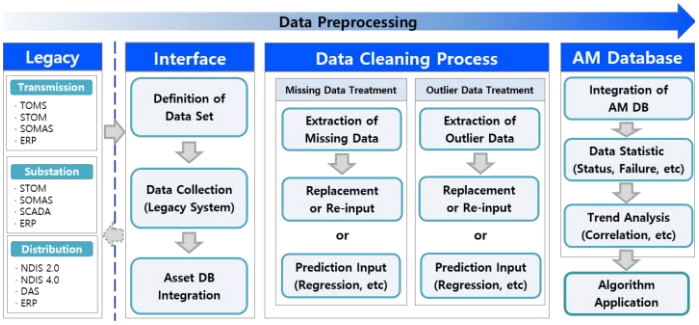

Fig. 2. Concept of asset data preprocessing.

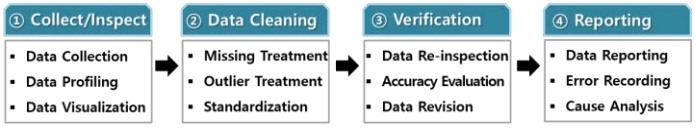

Fig. 3. Processing of asset data cleaning.

지식기반의 규칙이 필요하기 때문에 지중송전케이블 설비특성에 대해 지식베이스가 있는 전문가가 정제방법을 직접 구현할 수밖에 없다. 데이터 정제시간이 매우 길고 업무에 지장을 줄 뿐만 아니라 데이터가 시간에 따라 지속적으로 축적되기 때문에 정제는 지속적으로 이루어져야 한다.

따라서 본 논문에서는 지중송전케이블에 특화된 자산데이터 자동 정제 알고리즘과 그 알고리즘이 탑재된 시스템 개발 연구를 수행하였다. 정제 알고리즘은 결측데이터와 이상데이터를 처리하는 알고리즘에 대해 상세히 다루었다. 데이터 처리는 각각 규칙기반과 전문가 의견이 복합적으로 고려된 정제 알고리즘을 구현하였으며, 품질평가를 통해 정제 전후 변화를 확인하였다.

II. Process of Asset Data Cleaning

정제되지 않은 데이터를 오염데이터라고 하고, 오염데이터를 정상화시키는 것을 정제라고 한다. 데이터 중 오염데이터 비율이 일정 수준 이상이면, 통계 분석 결과가 무의미해질 뿐만 아니라 이러한 데이터를 사용하는 알고리즘의 신뢰성을 좌우할 수 있기 때문에 데이터의 정제과정이 반드시 필요하다. 또한, 빅데이터나 인공지능과 관련된 기술의 성능은 분석시 사용하는 데이터의 신뢰성에 달려있으며, 이를 위해 데이터 전처리를 통해 품질향상을 위한 핵심기술이 중요시되고 있다. 데이터 전처리란 Fig. 2와 같이 데이터 수집, 정제, 분석, 시각화 순서의 데이터 일련 업무이다. 데이터 업무에 소요되는 시간을 분석한 연구에 따르면, 데이터 정제가 60%를, 데이터 수집이 20%로써 대부분을 차지한다 [6]. 여기서 정제 프로세스는 Fig. 3과 같이 데이터 수집, 정제, 검증, 보고로 구성된다.

TABLE 1
Technology Comparison Between Conventional and Novel Automatic Cleaning Method

Item	Conventional Tech.	Automatic Cleaning Tech.
Cleaning Method	Manual cleaning by human	Automatic cleaning by algorithm
Cleaning Period	6 months/asset (approx.)	1 week/asset (approx.)
Data Quality	About 70%	Over 95%
Management Convenience	Uncomfortable	Comfortable

정제 과정은 수집 데이터에서 결측데이터를 찾아 처리한 후 이상데이터를 처리하는 순서로 이루어진다 [7]. 결측데이터란 데이터가 존재해야 하는 데이터이지만, 미입력에 의해 결측된 데이터를 의미한다. 이상데이터는 데이터가 존재하지만, 이상으로 의심되는 데이터이다. 정제 과정에서 결측데이터와 이상데이터를 전체데이터에서 분류 및 추출하는 게 우선시되어야 하며, 추출된 데이터를 평균값, 최빈값, 중간값 등의 대체값을 적용하거나 회귀분석을 통해 예측값을 적용하여 자동으로 정제시키는 방법이 일반적이다. 대용량의 빅데이터라면 일반적인 정제방법을 사용하더라도 오류데이터가 전체 중에 무시할 만한 일부라서 전체 빅데이터를 분석하는데 크게 지장을 받지 않을 수 있지만, 데이터의 양이 한정적이라면 잘못된 데이터로 정제 시 큰 영향을 받을 수 있기 때문에 설비 전문가가 데이터 재검사를 통해 정확도를 향상시키는 것이 중요하다.

정제 알고리즘의 장점은 데이터의 품질향상이외에도 시간 단축이 있다. 자산의 개체수가 많고, 점검 및 진단데이터가 많을수록 정제과정의 소요시간은 비례하여 증가하게 된다. 수집된 모든 데이터에서 수치가 이상하거나 기입이 되지 않았거나 하는 데이터를 분류하는 작업은 기존에는 사람이 수십만부터 수백만개의 데이터를 일일이 육안으로 확인하고 정제하였으나, 이는 데이터의 양이 많고 복잡하기 때문에 매우 오랜 시간이 소모되며, 인력을 낭비하게 되고 사람에 의해서 진행되다 보니 정제과정 중에 실수가 발생할 수 있다. 따라서 자동정제 기술이 개발되면 TABLE 1과 같이 종래기술 대비 정제기간이 1주일 내로 대폭 단축되고 데이터 정확도는 95% 이상으로 향상될 수 있을 것으로 사료된다.

또한, 새로운 데이터가 생성될 때마다 데이터를 수동정제해야 한다는 점과 다른 전력설비의 자산관리 확대 적용시 동일한 인적 정제작업을 수행해야 한다는 부담이 있기 때문에 자동정제 기술 적용시 관리측면에서 편리하다. 이러한 이유들로 인해 전력설비에 특화된 자산데이터 자동정제 알고리즘의 개발이 필요하지만, 상용 정제 소프트웨어는 광범위한 산업분야에서 적용되고 있고 [8]-[11], 전력회사의 설비자산 데이터에 대한 정제방법은 전 세계적으로 관련된 문헌이나 시스템이 거의 없다.

Item	Circuit Name	Cable Segment Code	Joint (start)	Joint (end)	Phase	Single /Double	Manu-facturer	Date of Installation
Raw Data	A-B #1	CTD0000001	J/B #1	J/B #2	A	S	OO전선	2012.01.01
	A-B #1	CTD0000002	J/B #1	J/B #2	B	S	OO전선	2012.01.01
	A-B #1	CTD0000003	J/B #1	J/B #2		S	OO전선	2012.01.01
Clean Data	A-B #1	CTD0000001	J/B #1	J/B #2	A	S	OO전선	2012.01.01
	A-B #1	CTD0000002	J/B #1	J/B #2	B	S	OO전선	2012.01.01
	A-B #1	CTD0000003	J/B #1	J/B #2	C	S	OO전선	2012.01.01

Fig. 4. Missing data cleaning algorithm example for phase information.

Item	Circuit Name	Cable Segment Code	Joint (start)	Joint (end)	Phase	Single /Double	Manu-facturer	Date of Installation
Raw Data	A-B #1	CTD0000001	J/B #1	J/B #2	A	S	OO전선	2012.01.01
	A-B #1	CTD0000002	J/B #1	J/B #2	B	S	OO전선	2012.01.01
	A-B #1	CTD0000003	J/B #1	J/B #2	C	S		2012.01.01
Clean Data	A-B #1	CTD0000001	J/B #1	J/B #2	A	S	OO전선	2012.01.01
	A-B #1	CTD0000002	J/B #1	J/B #2	B	S	OO전선	2012.01.01
	A-B #1	CTD0000003	J/B #1	J/B #2	C	S	OO전선	2012.01.01

Fig. 5. Missing data cleaning algorithm example for manufacturer.

III. Algorithm of Asset Data Cleaning

지중송전케이블 자산데이터 자동정제 방법 및 그 시스템은 결측데이터 처리 알고리즘과 이상데이터 처리 알고리즘으로 구성되어 알고리즘이 동작하며, 향후 자동정제 시스템에 알고리즘이 탑재되어 구현될 수 있다. 정제 알고리즘은 다양한 데이터 항목 중 대표적인 항목만 다루어 알고리즘 예시를 다루었다.

A. 결측데이터 처리 알고리즘

1) 규칙기반 결측데이터 처리

규칙기반 결측데이터는 데이터의 정보에 정답이 정해져 있는 설비정보에 주로 적용할 수 있다.

a) 케이블 상 정보

전력설비는 3상 시스템으로 A상, B상, C상으로 구성되어 항상 A상, B상, C상의 개수는 서로 일치해야 하는데 Legacy System에 오기입력하게 되면 일부 상 개수가 많거나 적은 오류가 발생할 수 있다. A, B, C상 순으로 입력되는 게 정상인데 Fig. 4와 같이 누락으로 A상, B상, C상 중 A상, B상만 입력하여 상 개수가 일치하지 않는 것을 확인할 수 있다. 이는 A상, B상, C상으로 입력하는 규칙이 존재하기 때문에 이를 이용하여 기존 "A상, B상, 누락"을 "A상, B상, C상"으로 정제할 수 있다.

b) 케이블 제작사 정보

케이블 제작사명은 수기 입력되는 경우가 많고 그 결과, 사람의 기호에 따라 입력되어 동일한 제작임에도 불구하고 명칭이 매우 다양하며 오탈자가 존재하고 있다.

대표적인 사례는 Fig. 5와 같이 "OO전선, OO전선, 누락"으로 입력된 부분을 "OO전선, OO전선, OO전선"으로 입력될 수 있다. 일반적으로 설치년월이 일치한다면 A상, B상, C상을 일괄 설치한 경우이기 때문에 한 상의 제작사 정보가 누락될지라도 다른 상의 제작사 정보를 통해 정제 가능하다.

Item	Circuit Name	Cable Segment Code	Joint (start)	Joint (end)	Phase	Single /Double	Manu-facturer	Date of Installation
Raw Data	A-B #1	CTD0000001	J/B #1	J/B #2	A	D1	OO전선	2012.01.01
	A-B #1	CTD0000002	J/B #1	J/B #2	A	D2	OO전선	2012.01.01
	A-B #1	CTD0000003	J/B #1	J/B #2	B	D1	OO전선	2012.01.01
	A-B #1	CTD0000004	J/B #1	J/B #2	B	D2	OO전선	2012.01.01
	A-B #1	CTD0000005	J/B #1	J/B #2	C	D1	OO전선	2012.01.01
	A-B #1	CTD0000006	J/B #1	J/B #2	C		OO전선	2012.01.01
Clean Data	A-B #1	CTD0000001	J/B #1	J/B #2	A	D1	OO전선	2012.01.01
	A-B #1	CTD0000002	J/B #1	J/B #2	A	D2	OO전선	2012.01.01
	A-B #1	CTD0000003	J/B #1	J/B #2	B	D1	OO전선	2012.01.01
	A-B #1	CTD0000004	J/B #1	J/B #2	B	D2	OO전선	2012.01.01
	A-B #1	CTD0000005	J/B #1	J/B #2	C	D1	OO전선	2012.01.01
	A-B #1	CTD0000006	J/B #1	J/B #2	C	D2	OO전선	2012.01.01

Fig. 6. Missing data cleaning algorithm example for single/double information.

Item	Circuit Name	Cable Segment Code	Joint (start)	Joint (end)	Phase	Single /Double	Manu-facturer	Date of Installation
Raw Data	A-B #1	CTD0000001	J/B #1	J/B #2	A	D1	OO전선	2012.01.01
	A-B #1	CTD0000002	J/B #1	J/B #2	A	D2	OO전선	
	A-B #1	CTD0000003	J/B #1	J/B #2	B	D1	OO전선	
	A-B #1	CTD0000004	J/B #1	J/B #2	B	D2	OO전선	2012.01.01
	A-B #1	CTD0000005	J/B #1	J/B #2	C	D1	OO전선	2012.01.01
	A-B #1	CTD0000006	J/B #1	J/B #2	C	D2	OO전선	2012.01.01
Clean Data	A-B #1	CTD0000001	J/B #1	J/B #2	A	D1	OO전선	2012.01.01
	A-B #1	CTD0000002	J/B #1	J/B #2	A	D2	OO전선	2012.01.01
	A-B #1	CTD0000003	J/B #1	J/B #2	B	D1	OO전선	2012.01.01
	A-B #1	CTD0000004	J/B #1	J/B #2	B	D2	OO전선	2012.01.01
	A-B #1	CTD0000005	J/B #1	J/B #2	C	D1	OO전선	2012.01.01
	A-B #1	CTD0000006	J/B #1	J/B #2	C	D2	OO전선	2012.01.01

Fig. 7. Missing data cleaning algorithm example for installation date information.

c) 케이블 S(Single)/D(Double) 정보

A변전소와 B변전소간 연계하는 케이블에 있어 송전용량이 많이 요구되는 경우 기존의 1개 케이블(Single) 대신 2개 케이블(Double)로 연계하는 경우가 있다. 이 때 Double인 경우 번들로 D1, D2로 구분하게 되며, D1과 D2는 하나의 조를 이루고 있으므로 D1의 개수와 D2의 개수가 일치해야 하나, 데이터 분석 시 개수가 서로 불일치 되는 경우를 Fig. 6과 같이 확인할 수 있다. 이는 상 정보와 함께 고려하여 A, A, B, B, C, C상 순으로 입력되었는지를 확인하고 기존 "D1, D2, D1, D2, D1, 누락"을 "D1, D2, D1, D2, D1, D2"로 규칙기반의 자동 정제가 가능하다.

d) 케이블 설치일자 정보

케이블 설치일자는 케이블의 운영년수(age)를 알 수 있어 설비 자산관리에 있어 매우 중요한 데이터이다. Legacy System에 수기입력에 따라 설치일자가 대다수는 2012년 1월에 설치되었는데 일부 설비의 설치일자 데이터 정보가 누락되었음을 알 수 있다. 케이블 회선은 일반적으로 장거리 선로이고 설치일자 데이터가 구간별로 존재하기 때문에 Fig. 7과 같이 설치일자 최빈값이 2012년 1월이라는 것을 알 수 있으므로 설치일자 데이터는 전체적인 패턴을 분석하여 정제가 가능하다는 것을 의미한다. 따라서 규칙을 부여하여 자동 정제할 수 있다.

2) 전문가 의견기반 결측데이터 처리

전문가 의견기반 자동정제 알고리즘은 데이터의 정보에 정답

Item	Before Change	Application Date	After Change
Termination (EBA)	▪ 154kV XLPE Cables - Insulator: Porcelain	2005.04.01.	▪ 154kV XLPE Cables - Insulator: Polymer

Fig. 8. Missing data cleaning algorithm example for termination insulator type information.

Item	Before Change	Application Date	After Change
Joint Box	▪ 154kV XLPE Cables - Joint method: TMJ	2006.03.01.	▪ 154kV XLPE Calbes - Joint method: PMJ

Fig. 9. Missing data cleaning algorithm example for jointing method type information.

이 정해지지 않은 다수의 결측데이터 처리에 주로 적용할 수 있다. 케이블 자산데이터 중 일부 데이터 항목은 결측률은 매우 높은 상태인데 이러한 결측률을 최소화시키기 위해서는 데이터 기입의 일괄 적용하는 방식을 채택할 수 있다.

a) EBA 종단접속함 애관종류 정보

지중송전 XLPE 케이블의 경우, 종단접속함(EBA)의 애관종류 또한 정보를 일일이 현장 확인하여 처리하기 곤란하고 장기간 소요가 예상되므로 Fig. 8과 같이 적용시점을 분석하여 그 시점 이전에는 자기애관, 이후에는 폴리머애관으로 자동 정제하고 적용시점의 ±1년을 확인하여 최종정제를 완료할 수 있다.

b) 중간접속함 접속공법 정보

지중송전 XLPE 케이블의 경우, 중간접속함의 접속공법은 초기에는 TMJ를 사용해오다가 PMJ 공법이 개발되면서 현재까지 PMJ를 사용하고 있다. 이 정보 또한 Fig. 9와 같이 적용시점을 알 수 있다면 자동 정제하고 적용시점의 ±1년을 확인하여 최종정제를 완료할 수 있다.

B. 이상데이터 처리 알고리즘

1) 규칙기반 이상데이터 처리

a) 케이블 상 정보

전력설비는 3상 시스템으로 A상, B상, C상으로 구성되어 항상 A상, B상, C상의 개수는 서로 일치해야 하는데 Legacy System에 오기입력하게 되면 일부 상 개수가 많거나 적은 오류가 발생할 수 있다. A, B, C상 순으로 입력되는 게 정상인데 Fig. 10과 같이 오기입으로 인해 A, B, B상으로 입력하여 상 개수가 일치하지 않는 것을 확인하고 상 입력 순서 패턴을 이용하여 불일치한 정보를 확인하여 정제할 수 있다. 만약 데이터를 정제하지 않고 방치할 경우, A-B #1 회선의 중간접속함 1번과 2번을 연결하는 회선의 B상은 2개로 오인될 수 있고, 이러한 데이터로 빅데이터 분석을 수행하면 전혀 다른 결과가 도출될 수 있다.

b) 케이블 제작사 정보

케이블 제작사명은 수기 입력되는 경우가 많고 그 결과, 사람

Item	Circuit Name	Cable Segment Code	Joint (start)	Joint (end)	Phase	Single /Double	Manu-facturer	Date of Installation
Raw Data	A-B #1	CTD0000001	J/B #1	J/B #2	A	S	OO전선	2012.01.01
	A-B #1	CTD0000002	J/B #1	J/B #2	B	S	OO전선	2012.01.01
	A-B #1	CTD0000003	J/B #1	J/B #2	B	S	OO전선	2012.01.01
Clean Data	A-B #1	CTD0000001	J/B #1	J/B #2	A	S	OO전선	2012.01.01
	A-B #1	CTD0000002	J/B #1	J/B #2	B	S	OO전선	2012.01.01
	A-B #1	CTD0000003	J/B #1	J/B #2	C	S	OO전선	2012.01.01

Fig. 10. Outlier data cleaning algorithm example for phase information.

Item	Circuit Name	Cable Segment Code	Joint (start)	Joint (end)	Phase	Single /Double	Manu-facturer	Date of Installation
Raw Data	A-B #1	CTD0000001	J/B #1	J/B #2	A	S	OO전선	2012.01.01
	A-B #1	CTD0000002	J/B #1	J/B #2	B	S	㈜OO전선	2012.01.01
	A-B #1	CTD0000003	J/B #1	J/B #2	C	S	OO전선㈜	2012.01.01
Clean Data	A-B #1	CTD0000001	J/B #1	J/B #2	A	S	OO전선	2012.01.01
	A-B #1	CTD0000002	J/B #1	J/B #2	B	S	OO전선	2012.01.01
	A-B #1	CTD0000003	J/B #1	J/B #2	C	S	OO전선	2012.01.01

Fig. 11. Outlier data cleaning algorithm example for manufacturer.

Item	Circuit Name	Cable Segment Code	Joint (start)	Joint (end)	Phase	Single /Double	Manu-facturer	Date of Installation
Raw Data	A-B #1	CTD0000001	J/B #1	J/B #2	A	D1	OO전선	2012.01.01
	A-B #1	CTD0000002	J/B #1	J/B #2	A	D2	OO전선	2012.01.01
	A-B #1	CTD0000003	J/B #1	J/B #2	B	D1	OO전선	2012.01.01
	A-B #1	CTD0000004	J/B #1	J/B #2	B	D2	OO전선	2012.01.01
	A-B #1	CTD0000005	J/B #1	J/B #2	C	D1	OO전선	2012.01.01
	A-B #1	CTD0000006	J/B #1	J/B #2	C	D1	OO전선	2012.01.01
Clean Data	A-B #1	CTD0000001	J/B #1	J/B #2	A	D1	OO전선	2012.01.01
	A-B #1	CTD0000002	J/B #1	J/B #2	A	D2	OO전선	2012.01.01
	A-B #1	CTD0000003	J/B #1	J/B #2	B	D1	OO전선	2012.01.01
	A-B #1	CTD0000004	J/B #1	J/B #2	B	D2	OO전선	2012.01.01
	A-B #1	CTD0000005	J/B #1	J/B #2	C	D1	OO전선	2012.01.01
	A-B #1	CTD0000006	J/B #1	J/B #2	C	D2	OO전선	2012.01.01

Fig. 12. Outlier data cleaning algorithm example for single/double information.

의 기호에 따라 입력되어 동일한 제작임에도 불구하고 명칭이 매우 다양하며 오탈자가 존재하고 있다. 대표적인 사례는 Fig. 11과 같이 OO전선으로 입력되어야 할 제작사명이 (주)OO전선, OO전선(주), OO전기, OO케이블과 같이 입력될 수 있다. 특정 제작사의 납품실적 조회, 고장데이터를 통한 통계적 수명분석 등 다양하게 사용 가능한 데이터가 미정제시 1개 제작사임에도 데이터상 2, 3개 제작사가 존재하는 것처럼 보이며, 1개 제작사의 데이터 추출 시 누락될 수 있기 때문에 제작사 명칭에 대한 통일화가 필요하다. 제작사명에 대한 수기입력 방식을 분석한다면 특정 패턴이 도출되므로 자동정제 방법을 구현할 수 있다.

c) 케이블 S(Single)/D(Double) 정보

A변전소와 B변전소간 Double 케이블로 연계하는 경우가 있다. 이 때 Double인 경우 번들로 D1, D2로 구분하게 되며, D1과 D2는 하나의 조를 이루고 있으므로 D1의 개수와 D2의 개수가 일치해야 하나, 데이터 분석 시 개수가 서로 불일치되는 경우를 확인할 수 있다. 이는 상 정보와 함께 고려하여 A, A, B, B, C, C 순으로 입력되었는지를 확인하고 D1, D2, D1, D2, D1, D2 순의 패턴으로 입

Item	Circuit Name	Cable Segment Code	Joint (start)	Joint (end)	Phase	Single /Double	Manu-facturer	Date of Installation
Raw Data	A-B #1	CTD0000001	J/B #1	J/B #2	A	D1	OO전선	2012.01.01
	A-B #1	CTD0000002	J/B #1	J/B #2	A	D2	OO전선	2021.01.01
	A-B #1	CTD0000003	J/B #1	J/B #2	B	D1	OO전선	0212.01.01
	A-B #1	CTD0000004	J/B #1	J/B #2	B	D2	OO전선	2012.01.01
	A-B #1	CTD0000005	J/B #1	J/B #2	C	D1	OO전선	2012.01.01
	A-B #1	CTD0000006	J/B #1	J/B #2	C	D2	OO전선	2012.01.01
Clean Data	A-B #1	CTD0000001	J/B #1	J/B #2	A	D1	OO전선	2012.01.01
	A-B #1	CTD0000002	J/B #1	J/B #2	A	D2	OO전선	2012.01.01
	A-B #1	CTD0000003	J/B #1	J/B #2	B	D1	OO전선	2012.01.01
	A-B #1	CTD0000004	J/B #1	J/B #2	B	D2	OO전선	2012.01.01
	A-B #1	CTD0000005	J/B #1	J/B #2	C	D1	OO전선	2012.01.01
	A-B #1	CTD0000006	J/B #1	J/B #2	C	D2	OO전선	2012.01.01

Fig. 13. Outlier data cleaning algorithm example for installation date information.

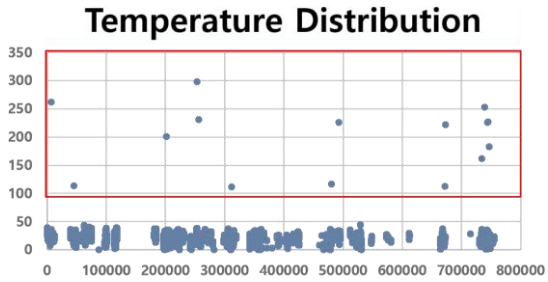

Fig. 14. Outlier data cleaning algorithm example for single/double information.

Item	Circuit Name	Joint Box Name	Joint Type	Current	Phase	Single /Double	Ambient Temperature	Measured Temperature
Raw Data	A-B #1	J/B #1	IJ	217[A]	A	S	20℃	22℃
	A-B #1	J/B #1	IJ	217[A]	B	S	20℃	22℃
	A-B #1	J/B #1	IJ	217[A]	C	S	20℃	222℃
Clean Data	A-B #1	J/B #1	IJ	217[A]	A	S	20℃	22℃
	A-B #1	J/B #1	IJ	217[A]	B	S	20℃	22℃
	A-B #1	J/B #1	IJ	217[A]	C	S	20℃	22℃

Fig. 15. Outlier data cleaning algorithm example for measured temperature.

력되었는지를 동시 고려하여 패턴 불일치 사항을 확인하고 Fig. 12와 같이 정제 가능하다.

d) 케이블 설치일자 정보

케이블 설치년월은 케이블의 운영년수를 알 수 있는 매우 중요한 데이터임에도 불구하고 수기입력에 따라 설치년월이 Fig. 13과 같이 2012년 1월에 설치됨에도 불구하고, 아직 도래하지 않은 2021년 1월이라든지, 의미를 알 수 없는 0212년 1월로 입력되는 경우가 존재한다. 케이블 회선은 일반적으로 장거리 선로이고 설치년월 데이터가 구간별로 존재하기 때문에 전반적인 설치년월이 2012년 1월이라는 것을 알 수 있으므로 이는 설치년월 데이터는 전체적인 패턴을 분석하여 정제가 가능하다는 것을 의미한다. 따라서 현재 일자 기준으로 도래하지 않은 설치년월이나, 알 수 없는 데이터는 앞뒤 정보에서 가장 많이 나온 데이터를 적용한다든지 등의 규칙을 부여하여 자동 정제할 수 있다.

2) 전문가 의견기반 이상데이터 처리

a) 케이블 및 접속함 열화상 정보

케이블 및 접속함 점검 시 열화상카메라를 이용하여 최대온도를 측정하여 기입하는데, 수기입력 시 오기입이 발생하여 정제할 필요가 있다. 일례로 최대온도 22°C로 측정된 값을 기입 시 22°C, 22°C 순으로 입력하다 오타로 222°C를 잘못 입력되어 있는 경우 데이터 프로파일링 기법을 사용하여 Fig. 14와 같이 적정 온도범위 외 이상데이터만 추출하고, 이상값 처리방법으로 정제한다. 222°C 값은 실제로 발생하기 힘든 값으로써 추출되면 다른 상에 어떻게 입력되었는지 확인 후 통계적으로 평균을 취해 Fig. 15와 같이 22°C로 대체 가능하다.

케이블 및 접속함 열화상 정보 이외에도 OF 케이블 절연유의 유중가스분석, XLPE 케이블 종단접속함의 가스, 절연유 분석에도 이상데이터 군집을 찾아 분석하여 정제가 가능하다. 한편, 데이터 항목마다 이상데이터 발생시 처리하는 방법이 각각 다를 수 있기 때문에 데이터 프로파일링, 통계적 방법, 선형 회귀 함수 등을 적용해보고 어떤 방법이 가장 적정한 지를 분석하여 정확성을 높일 수 있는 방식을 채택하여 알고리즘으로 구현할 수 있다.

따라서 알고리즘 구현 전 데이터 항목별로 조사가 필요하며, 최적화시 데이터 축적에도 지속적인 정제가 가능하다.

자산데이터 자동 정제 알고리즘이 탑재되어 프로그램화 시킨 것은 자산데이터 자동 정제 시스템이며, 전력설비의 현황/상태/점검/기타 등 데이터를 시스템에 입력하고 자동 정제 프로세스를 통해 정제된 순수한 데이터베이스를 출력할 수 있을 뿐만 아니라 원시데이터와의 비교를 통해 변경내역까지 확인할 수 있다. 또한, 데이터의 품질 분석 기능을 통해 정제 전후의 품질 변화를 볼 수 있어 어떤 데이터 항목의 품질이 저조한지를 포함하여 전체적인 품질 양상을 확인할 수 있다. 이와 같이 자산데이터 자동 정제 시스템은 개발 중이며, 시스템에 탑재될 기본 정제 알고리즘은 전술한 내용을 근간으로 한다. 시스템 개발 시 국내 최초로 전력설비 자산데이터에 대한 정제 기술이 정립 및 구현되고, 데이터 정확성을 95% 이상까지 향상시킬 수 있을 것으로 기대한다.

IV. Conclusion

본 논문에서는 지중송전케이블 자산데이터에 대한 자동정제 알고리즘에 대한상세한 예시를 들어 설명하였다. 정제 알고리즘은 결측데이터 처리와 이상데이터 처리로 구분되며 규칙기반과 전문가 의견기반의 정제방법을 융합하여 사용함으로써 기존 수기정제 방식과 차별성을 보여준다. 전력설비에 특화된 자동정제 기술은 세계적으로 아직까지 공개되지 않은 기술로써 알고리즘과 시스템 구현기술 확보 시 정제데이터 확보를 통해 통계분석이나 인공지능, 빅데이터 분석 등에 다방면으로 활용 가능할 것으로 예상된다. 또한, 자산관리시스템의 핵심 성능평가 알고리즘에 고품질의 정제데이터가 사용됨으로써 설비 교체 투자 우선순위 결과에 대한 최소한의 신뢰성 확보가 가능할 것으로 기대된다.

References

[1] ISO 55000, "Asset Management - Overview, Principles and

Terminology," 2014.

[2] ISO 55001, "Asset Management – Management Systems – Requirements," 2014.

[3] ISO 55002, "Asset Management – Management Systems - Guidelines for the Application of ISO 55001," 2014.

[4] FERF (Financial Education & Research Foundation), "Data: The Strategic Asset," 2019.

[5] IEC White Paper, Strategic Asset Management of Power Networks, 2015.

[6] CrowdFlower, "Data Science Report," 2016.

[7] S. Burke, "Missing Values, Outliers, Robust Statistics & Non-parametric Methods," LC GC Europe, 2001.

[8] https://openrefine.org.

[9] https://www.trifacta.com.

[10] https://winpure.com.

[11] https://www.paxata.com.