JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 12, pp. 129-136, December 2023
https://doi.org/10.9708/jksci.2023.28.12.129

A Study on the Domain Discrimination Model of CSV Format Public Open Data

Ha-Na Jeong*, Jae-Woong Kim**, Young-Suk Chung*

*Student, Dept. of Computer Engineering, Kongju National University, Cheonan, Korea
**Professor, Dept. of Software Engineering, Kongju National University, Cheonan, Korea
*Lecturer, Dept. of Computer Engineering, Kongju National University, Cheonan, Korea

[Abstract]

The government of the Republic of Korea is conducting quality management of public open data by conducting a public data quality management level evaluation. Public open data is provided in various open formats such as XML, JSON, and CSV, with CSV format accounting for the majority. When diagnosing the quality of public open data in CSV format, the quality diagnosis manager determines and diagnoses the domain for each field based on the field name and data within the field of the public open data file. However, it takes a lot of time because quality diagnosis is performed on large amounts of open data files. Additionally, in the case of fields whose meaning is difficult to understand, the accuracy of quality diagnosis is affected by the quality diagnosis person's ability to understand the data. This paper proposes a domain discrimination model for public open data in CSV format using field names and data distribution statistics to ensure consistency and accuracy so that quality diagnosis results are not influenced by the capabilities of the quality diagnosis person in charge, and to support shortening of diagnosis time. As a result of applying the model in this paper, the correct answer rate was about 77%, which is 2.8% higher than the file format open data diagnostic tool provided by the Ministry of Public Administration and Security. Through this, we expect to be able to improve accuracy when applying the proposed model to diagnosing and evaluating the quality management level of public data.

▸ **Key words**: Open data, Data quality, Quality improvement, Data quality diagnosis, Data Distribution

• First Author: Ha-Na Jeong, Corresponding Author: Jae-Woong Kim
*Ha-Na Jeong (konghanaj@gmail.com), Dept. of Computer Engineering, Kongju National University
**Jae-Woong Kim (jykim@kongju.ac.kr), Dept. of Software Engineering, Kongju National University
*Young-Suk Chung (merope@kongju.ac.kr), Dept. of Computer Engineering, Kongju National University
• Received: 2023. 10. 18, Revised: 2023. 12. 07, Accepted: 2023. 12. 08.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

정부는 공공데이터 품질관리 수준평가를 진행하여 공공 개방데이터의 품질관리를 진행하고 있다. 공공 개방데이터는 XML, JSON, CSV 등 여러 오픈포맷 형태로 제공되며 CSV 형식이 대다수를 차지한다. 이러한 CSV 형식의 공공 개방데이터 품질진단 시 품질진단 담당자가 공공 개방데이터 파일의 필드명과 필드 내 데이터에 의존하여 필드 별 도메인을 판단하여 진단한다. 그러나 대량의 개방 데이터 파일을 대상으로 품질진단을 수행하기 때문에 많은 시간이 소요된다. 또한 의미 파악이 어려운 필드의 경우 품질진단의 정확성이 품질진단 담당자의 데이터 이해도 역량의 영향을 받는다. 본 논문은 필드명과 데이터 분포 통계를 이용한 CSV 형식 공공 개방데이터의 도메인 판별 모델을 제안하여 품질진단 결과가 품질진단 담당자의 역량에 좌지우지 되지 않도록 일관성과 정확성을 보장하고 진단 소요 시간 단축을 지원한다. 본 논문의 모델 적용 결과 행정안전부에서 제공하는 파일형식 개방데이터 진단도구보다 2.8% 높은 약 77%의 정답률을 보였다. 이를 통해 공공데이터 품질관리 수준진단·평가에 제안 모델 적용 시 정확성을 향상시킬 수 있을 것으로 기대한다.

▸ **주제어**: 개방데이터, 데이터품질, 품질개선, 데이터 품질진단, 데이터 분포

I. Introduction

공공데이터란 정부가 생산하여 보유 및 관리하고 있는 데이터를 의미한다[1]. 공공데이터를 민간에 개방하여 누구든지 공공데이터를 활용하여 수익 창출이 가능하도록 개방하는 것이 오픈데이터 정책이다. 오픈데이터 정책은 전 세계적으로 주목받고 있다. 영국, 미국 등 많은 선진국에서 공공데이터 개방을 통해 데이터 유동화 및 민주화를 도모할 수 있다는 점을 들어 적극적으로 오픈데이터 정책을 추진하고 있다[1-3]. 정부는 공공데이터의 제공 및 이용 활성화에 관한 법률을 제정하고 이것을 기반으로 공공데이터 개방을 진행하고 있다[4]. 이러한 공공데이터 개방은 민간의 삶의 질 향상과 일자리 창출, 신산업 육성 등 국민경제 발전에 이바지하는 것을 목적으로 두고 있다[5-6]. 공공데이터는 공공데이터법에 제2조 제3호에 의거하여 기계판독이 가능한 형태로 제공되어야 하므로 공공데이터 포털(data.go.kr)을 통해 오픈포맷 형태의 데이터로 개방되고 있다[7].

다음의 Fig. 1.은 공공데이터 개방 증가 추세 그래프이다.

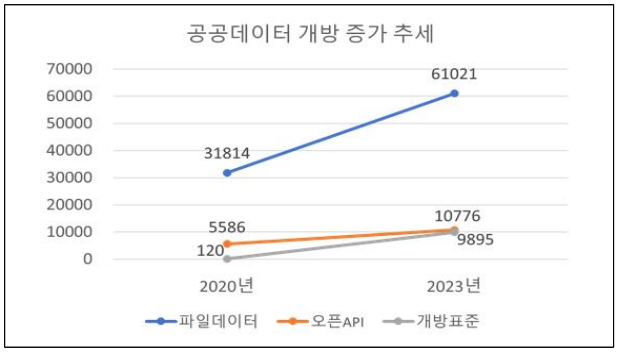

Fig. 1. Graph of Increasing Trend in Public Data Opening

2023년 8월 현재 공공데이터 포털에는 공공데이터가 오픈API 10,776건, 파일데이터 61,021건, 표준데이터셋 9,895건으로 총 81,692건 등록되어 있다[8]. 2020년도의 총 등록 건수인 37,520건과 비교하였을 때 3년 사이에 2배 이상 증가한 수치이다[9]. 그러나 이러한 공공데이터의 개방 증가 추세에 비해 공공데이터의 활용도는 기대에 미치지 못하고 있다. 그 주요한 원인으로는 공공데이터 품질관리와 표준화 미흡 등으로 인한 오류 데이터 혼재가 제기되고 있다[10-11]. 정부는 이를 해결하고자 한국지능정보화진흥원을 통해 공공데이터 품질관리 수준진단·평가를 실시하여 공공 개방데이터의 오류에 대한 진단 및 개선 작업을 수행하고 있다[12-13]. 데이터의 도메인 판별은 데이터 품질진단의 핵심 요소이다. 4차 산업혁명의 대두와 데이터의 활용도 증가로 인하여 빅데이터를 대상으로 한 도메인 판별 연구가 진행되고 있다[14]. 그러나 해당 연구들은 파일데이터가 아닌 데이터베이스를 대상으로 한다. 그러나 공공 개방데이터 81,692건 중 CSV형식으로 제공되고 있는 데이터는 51,104건으로 전체 중 약 62.6%를 차지하므로 CSV형식의 공공데이터를 대상으로 연구를 진행하였다. 데이터 필드명과 필드 내 데이터에 의존하여 수작업으로 데이터의 도메인을 판별해야 하는 특성을 가지는 공공 개방데이터의 오류진단은 시간과 비용을 많이 소비하고 품질진단의 정확성을 보장할 수 없다. 이를 해결하기 위해 본 논문에서는 CSV(Comma-Separated Values) 형식 공공 개방데이터의 도메인 판별 모델에 관한 연구를 진행하였다. 제안한 모델을 활용하면 데이터 진단에 소비되는

시간을 단축할 수 있고 진단 결과의 정확성과 일관성을 일정 수준으로 유지할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구와 관련된 데이터 품질진단 기준과 공통표준사전에 대해 서술한다. 3장에서는 본 연구의 모델을 제안한다. 4장에서는 실제 데이터에 제안된 모델에 적용하여 결과를 확인한다. 마지막 5장에서 본 연구의 결과를 정리하고 향후 연구과제에 대해 논의한다.

II. Preliminaries

1. Backgrounds

1.1 Data quality diagnosis standards

행정안전부와 한국지능정보사회진흥원이 주관하여 매년 진행되는 공공데이터 품질관리 수준진단·평가 사업은 '공공데이터의 제공 및 이용 활성화에 관한 법률' 제22조 공공데이터 품질관리에 근거를 두고, 공공기관이 생성 또는 취득하여 관리하는 공공데이터의 품질을 확보하기 위해 체계적으로 품질관리 활동을 수행하는가에 대한 여부를 진단하기 위한 목적으로 추진되고 있다[15]. 이 공공데이터 품질관리 수준진단·평가 사업은 진단 항목 및 지표, 절차 및 일정 등을 안내하기 위해 '공공데이터 품질관리 수준진단·평가 매뉴얼'을 배포하고 있다. 다음의 Table 1.은 2022년 공공데이터 품질관리 수준진단·평가 매뉴얼에 기재되어 있는 데이터 품질진단 기준 중 도메인 관련 항목을 정리한 것이다[16].

Table 1. List of quality diagnosis criteria for public data quality management level diagnosis evaluation

Type of diagnosis	Explanation	DB Y/N	File Y/N
Date domain	Measurement of data errors where date data values are out of valid range or format is out of standard	Y	Y
Number domain	Error measurements when number generation rules are violated	Y	Y
Whether domain	Measurement of data errors outside the range of valid values	Y	Y
Code domain	Measurement of data errors other than code values defined by the source DB as standard	Y	N
Amount domain	Measurement of data errors that contain characters other than amount data	Y	Y
Quantity domain	Measurement of data errors with characters other than quantity data	Y	Y
Percentage domain	Measurement of data errors that contain non-rate characters	○	○

Table 1.의 'DB Y/N' ,'File Y/N' 항목은 데이터베이스 진단 여부, 파일데이터 진단 여부를 뜻한다. 7종의 도메인 중 코드 도메인은 파일진단 대상에서 제외된다. 이는 코드 도메인의 경우 진단을 위해 코드값이 필요하기 때문이다. 파일데이터의 경우 코드값을 별도로 제공하지 않기 때문에 진단할 수 없으므로 진단대상에서 제외된다. 행정안전부에서 제공하는 공통표준용어, 공통표준도메인 또한 이 7종의 도메인으로 분류되어 제공되고 있다.

1.2 Common Standard Dictionary

행정안전부는 '공공데이터의 제공 및 이용 활성화에 관한 법률' 제23조에 따라 공공데이터를 누구나 동일한 의미로 이해, 사용할 수 있도록 하기 위해 공통표준사전을 정의하였다. '공공기관의 데이터베이스 표준화 지침' 제 8조에 의거하여 공공기관은 신규 데이터베이스를 구축할 경우 공통표준용어를 적용해야 하며 운영 중인 데이터베이스의 경우 신규 전면 구축/재구축 업무를 추진할 경우 이를 적용해야한다[17]. 행정안전부는 공통표준사전을 주기적으로 검토하여 공통표준사전을 추가 제정하거나 일부 폐기하는 등의 개선작업을 꾸준히 진행하고 있다.

공통표준사전은 공통표준용어, 공통표준도메인, 공통표준단어로 구성된다. 이 중 공통표준도메인은 도메인 진단기준과 유사하게 분류되어있으며 분류항목을 '공통표준도메인그룹명'으로 표기하고 있다, 이 '공통표준도메인그룹명' 은 상세분류 값을 가지는데 이것을 '공통표준도메인분류명'이라고 표기하고 있다.

다음의 Table 2.는 5차 제정 공통표준도메인의 공통표준도메인그룹명과 공통표준도메인분류명의 목록이다[18].

Table 2. List of classification names by group name of the 5th established common standard domain

Domain group name	Domain classification name
Amount	Price, Amount ···
Date/Time	Hour minute, Year ···
Contents	Contents
Designation	Name, Address ···
Number	Resident registration number, Business registration number ···
Quantity	latitude, longitude ···
Percentage	Ratio
Code	Whether, Code ···

데이터 품질 진단기준이 도메인 진단기준이 7종인 것과는 달리 공통표준도메인그룹은 8종으로 구성되어 있다. 이것은 공통표준도메인그룹 중 '내용(Contents)' 그룹은 진단이 불가능한 항목이기 때문이다. '내용' 그룹에 속하는 데이터는 게시판 글 등의 규칙성이 없는 단순 문자열이기 때문에 진단을 하지 않는다. 도메인 진단기준과 공통표준도메인그룹은 '날짜 도메인-날짜/시간', '번호 도메인-번호', '여부 도메인-코드(여부)', '금액 도메인-금액', '수량 도메인-수량', '율 도메인-율' 로 매핑된다.

III. The Proposed Scheme

1. Proposed model diagram

공개되는 공공데이터가 파일 형태로 되어 있어 파일데이터를 대상으로 한 연구가 필요하다. CSV 형식의 공공 개방데이터 품질진단 시 품질진단 담당자는 CSV 파일의 각 필드별로 진단규칙을 설정해야한다. 진단규칙을 설정하기 위해선 각 필드의 도메인을 파악해야한다. 그러나 품질진단 담당자가 해당 공공 개방데이터에 대한 이해도가 부족하거나 CSV 파일 내 필드명이 데이터의 의미를 파악하기 어렵게 설정되어있는 경우 필드의 도메인 파악에 어려움이 있고 품질진단의 정확성이 담당자의 역량에 따라 좌지우지된다. 또한 대량의 CSV 파일을 파악 후 진단해야 하므로 많은 시간이 소요된다. 이러한 문제점을 해결하기 위해 본 논문에서 CSV 형식 공공 개방 데이터의 도메인 판별 모델을 제안한다. 해당 모델은 CSV 파일의 필드명과 필드 내 데이터의 데이터 분포 통계를 이용하여 필드의 도메인을 판별한다.

다음의 Fig. 2.는 본 논문에서 제안하는 CSV 형식 공공 개방데이터의 도메인 판별 모델의 시퀀스 다이어그램이다.

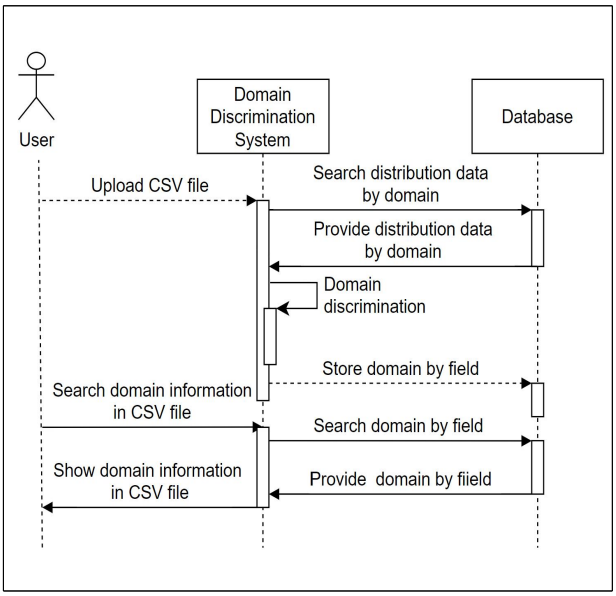

Fig. 2. Sequence Diagram of Domain Identification Model for CSV Format Public Open Data

CSV 형식 공공 개방데이터의 숫자데이터 도메인 판별 모델은 도메인 판별 시스템과 도메인 판별에 필요한 데이터를 저장하는 데이터베이스로 구성하였다. 데이터베이스에는 사전에 도메인별 분포 데이터가 저장되어 있고, 이를 CSV 형식의 공공 개방데이터 파일의 도메인을 판별할 때 활용한다. 사용자가 도메인 판별 시스템에 CSV 형식의 공공 개방데이터 파일을 업로드하면 도메인 판별 시스템이 CSV 파일의 각 필드별로 도메인별 분포 데이터 등 도메인 판별에 필요한 데이터를 데이터베이스에 조회한다. 데이터베이스에서 조회한 데이터를 이용하여 도메인을 판별하고, 판별된 필드별 도메인 정보를 데이터베이스에 저장한다. 그 후 사용자가 CSV 파일의 도메인 정보를 조회하면 도메인 판별 시스템은 데이터베이스에 저장되어 있는 필드별 도메인 정보를 사용자 측에 제공하는 구조이다.

2. Proposed model process

Fig. 3.은 본 논문에서 제안하는 CSV 형식 공공 개방데이터의 도메인 판별 모델의 도메인 판별 프로세스이다.

사용자가 CSV 파일 형식의 공공 개방데이터를 업로드하면 파일 내의 각 필드들의 도메인을 판별하기 위해 필드의 개수만큼 loop가 실행되며 각 필드의 도메인을 판별한다. loop내 프로세스의 단계별 상세 내용은 다음과 같다.

첫 번째, 필드의 필드명을 이용하여 1차적으로 후보 도메인을 추출한다. 이때 후보 도메인이 1개인 경우, 후보 도메인이 2개 이상 또는 0개인 경우에 따라서 프로세스가 분기된다.

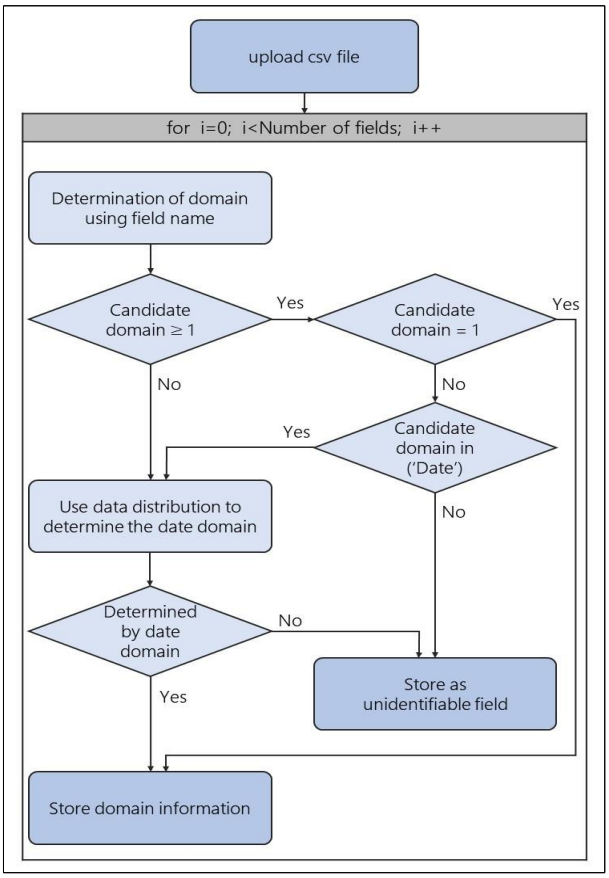

Fig. 3. Domain Determination Process

두 번째, 후보 도메인이 1개인 경우 후보 도메인을 해당 필드의 도메인으로 저장한다. 이때 CSV 파일명, 필드명, 필드 순서, 도메인 정보를 데이터베이스에 저장한다.

세 번째, 후보 도메인의 개수가 2개 이상이고 후보 도메인에 날짜 도메인이 포함되어있는 경우에는 데이터 분포를 이용한 날짜 도메인 판별을 진행하여 날짜 도메인으로 판별되는 데이터가 50% 이상이라면 날짜 도메인을 해당 필드의 도메인으로 저장한다. 예를 들어, 후보 도메인이 날짜 도메인, 수량 도메인 2가지일 때 필드 내 데이터가 날짜 도메인 분포를 이용하여 날짜 도메인 판별을 진행한다. 필드 내 데이터의 50% 이상이 날짜 도메인으로 판별된다면 해당 필드를 날짜 도메인으로 판별하여 저장한다.

마지막으로, 후보 도메인이 존재하지 않는 경우에는 3번째 프로세스와 동일하게 데이터 분포를 이용한 날짜 도메인 판별을 진행한다. 필드 내 데이터의 50% 이상이 날짜 도메인으로 판별된다면 해당 필드를 날짜 도메인으로 판별하여 저장한다.

첫 번째 프로세스에서 필드명을 이용하여 후보 도메인을 추출하는 방식은 다음과 같다. 다음 Table 3.은 도메인별 필드명 분류 기준이다.

Table 3. Field Name Classification Criteria by Domain

Domain	Classification Criteria Word
Date domain	Date, Time, Year, Month, Year-month-day, Start date, End date, **Day**
Number domain	Number, Main number, Sub-number, Contact, Phone call, Serial number
Whether domain	Whether, Existence and nonexistence
Amount domain	Amount, Publicly announced price, Balance, **Tax**, Unit price, Fee
Quantity domain	**Number**, Area, Size, Age, Latitude, Longitude, Quantity, Population
Percentage domain	**Rate**

도메인별 필드명 분류기준 단어는 5차 공통표준용어를 도메인 진단기준별로 분류 후 빈출 단어를 추출한 것, 공공데이터 제공표준의 빈출 단어를 추출한 것, 400개의 CSV 형식 공공 개방 데이터의 빈출 단어를 추출한 것을 정리하여 설정하였다[19]. 필드명에 특정 단어가 포함되어 있다면 해당 도메인을 후보 도메인으로 지정하는 방식이다. 단 굵은 글씨로 표기한 단어들은 필드명 문자열의 끝에 위치되어 있는 경우에 후보 도메인이 지정되도록 조건을 추가하였다. 또한 날짜 도메인 진단기준의 분류 기준 단어 중 '연도'의 경우 '년도' 문자열도 추가 허용하였다.

본 논문에서는 숫자 데이터의 경우에만 2차진단을 진행하기 때문에 숫자 외의 특수문자나 문자열이 포함된 데이터의 진단은 불가능하다. 그러나 날짜 도메인에 한하여 전처리 과정을 거쳐 단순 숫자 데이터로 변환하여 진단을 진행한다. 예를 들어, '연-월-일' 형식의 날짜 데이터를 특수문자를 제거한 '연월일' 형식의 데이터로 변환한다.

IV. Evaluation

공공데이터포털에서 제공하는 공공 개방데이터 중 300개의 CSV 형식의 데이터를 다운로드받아 본 논문에서 제안한 CSV 형식 공공 개방데이터의 도메인 판별 모델을 적용하여 결과를 도출하였다. 300개의 CSV 파일의 총 필드 수는 2,334개이며, 이 중 숫자형 데이터가 아닌 데이터, 기본 형식에서 벗어난 데이터들은 진단 불가능 필드로 분류되었다. 기본 형식에서 벗어난 데이터에는 '조합설립인가일자' 필드명 내에 '2008-09-24(2021-08-18)' 형식의 데이터가 들어간 것 등이 있다. 기본 형식을 만족하기 위해선 '2008-09-24' 형식과 같이 단일 값으로 이루어져 있어야 한다. 이러한 이유로 진단 대상에서 제외된 필드 수

는 1,212개이며 모델에 적용한 필드 수는 1,122개이다. 다음 Table 4.는 필드의 필드명을 이용하여 후보 도메인을 추출한 결과이다.

Table 4. List of Candidate Domain Extraction Results

Domain	Result Cnt	Actual Cnt	Correct Rate
Date domain	246	277	88%
Number domain	292	321	90%
Whether domain	6	8	73%
Amount domain	16	22	73%
Quantity domain	309	489	63%
Percentage domain	5	5	100%

Table 4.에서 'Result Cnt' 항목은 도메인 진단기준별 후보 도메인 개수이고 'Actual Cnt' 항목은 1,122개 각 필드들을 분석하여 실제로 속하는 도메인 진단기준별로 정리한 것이다. 율 도메인이 100%로 가장 높은 정답률을 보였고, 수량 도메인이 63%로 가장 낮은 정답률을 보였다. 진단 대상 필드 1,122개 중 874개의 필드가 후보 도메인이 1개 이상 추출되었다. 이 과정에서 후보 도메인이 실제 도메인과 다르게 설정된 필드들이 있다. 예를 들어 '월배출량'이라는 필드명의 경우 '월'이라는 날짜 도메인 필드명 분류 기준 단어를 포함하고 있어 날짜 도메인으로 오분류 되었다. 진단 대상 필드 1,122개 중 후보도메인이 옳게 분류된 것은 810개로 약 72.2%의 정답률을 보였다.

후보 도메인이 판별되지 않은 필드와 날짜 도메인이 포함된 2개 이상의 후보 도메인을 가진 필드에 대해 날짜 도메인 분포를 이용하여 날짜 도메인 판별을 진행하였다. 이때 도메인 분포는 공공데이터포털에서 제공하는 공공 개방데이터 중 400개의 CSV 형식 공공 개방데이터에서 추출하였다. 이 400개의 공공 개방데이터는 본 논문의 모델에 적용한 300개의 공공 개방데이터와 중복되지 않는다. 400개의 공공 개방데이터의 필드 데이터를 확인하여 수동으로 날짜 도메인에 해당하는 필드를 정리하였고 특수문자와 공백을 제거한 필드 데이터의 주 형식별로 데이터 분포를 분석하였다. 형식별로 데이터의 최대값, 최소값을 파악하고 구간을 나누어 구간별 데이터 빈도수를 추출하였다. 빈도수가 증가하는 구간을 지정하여 해당 구간에 속하는 데이터는 날짜 도메인에 속하는 데이터로 판별하였다. 다음의 Table 5.는 데이터 형식별 데이터 판별 범위이다.

Table 5. Data Determination Range by Data Format

Data Format	Discrimination Range
YYYY	2016 ~ 2023
YYYYMMDD	20080901 ~ 20230831
YYYYMMDDHHMISS	20190601000000 ~ 20230831115959

연월일 데이터인 "YYYYMMDD" 형식 데이터와 연월일 시분초인 "YYYYMMDDHHMISS" 형식 데이터의 경우 데이터 판별 정확도를 높이기 위하여 실 구간에서 일, 시, 분, 초 구간을 조정하여 판별 범위를 지정하였다. 예를 들어, "YYYYMMDD" 형식의 실제 판별 범위로 선정된 시작 구간 값은 "20080908"이지만 "20080901"로 조정하였다.

다음 Table 6.은 후보 도메인이 0개인 필드를 데이터 분포를 이용하여 날짜 도메인 판별을 진행한 결과이다.

Table 6. Date Domain Determination Progress Result

Data format	Real Cnt	Discrimination Cnt
YYYY	9	9
YYYYMMDD	54	42
YYYYMMDDHHMISS	4	3

후보 도메인이 0개인 510개의 필드 중 날짜 도메인에 속하지 않는 필드가 날짜 도메인으로 오분류된 경우는 발생하지 않았다. "YYYY" 형식 필드의 경우 정답률이 100%, "YYYYMMDD" 형식 필드의 경우 정답률이 약 78%, "YYYYMMDDHHMISS" 형식 필드의 경우 정답률이 75%이다. 즉, 데이터 분포를 이용한 날짜 도메인 판별 진행 결과 약 81%의 정답률을 보였다. 최종적으로 본 논문에서 제안한 모델을 적용하여 필드별 도메인을 판별한 결과 1,122개 필드 중 864개가 옳게 분류되어 77%의 정답률을 보였다.

V. Conclusions

본 연구는 CSV 형식 공공 개방데이터 품질진단 시 일관성, 정확성 보장과 진단 소요 시간 단축 지원을 위하여 CSV 형식 공공 개방데이터의 도메인 판별 모델을 제안하였다. 제안 모델은 필드명을 이용하여 CSV 형식 공공 개방데이터의 필드별 도메인을 1차 분류하고, 날짜 공공 개방데이터의 평균 데이터 분포를 산출하여 날짜 도메인을 2

차 분류한다. CSV의 필드명을 기본으로 분류하되, 필드명만으로 분류가 어려운 필드를 데이터 분포 통계를 이용하여 2차 분류 함으로써 도메인 분류의 정확성을 높이는 것이 CSV 형식 공공 개방데이터의 도메인 판별 모델의 특징이다. 모델을 적용한 결과 필드명으로 1차 분류하였을 때 약 72.2%, 데이터 분포 통계를 이용하여 날짜 도메인을 2차 분류하였을 때 약 77.0%의 정답률을 보였다. 이는 행정안전부에서 제공하는 파일형식 개방데이터 진단도구의 진단규칙 추천 기능의 정답률인 74.2%와 비교하였을 때 약 2.8% 높은 수치이다. 이를 통하여 제안 모델이 공공데이터 품질관리 수준진단·평가에 있어 유용한 모델임을 확인하였으며 공공데이터 품질관리 수준진단·평가에 제안 모델 적용 시 정확성을 향상시킬 수 있다. 제안한 모델을 활용하면 기존의 방식에 비하여 진단에 소비되는 시간을 단축할 수 있고 품질진단 담당자의 역량에 관계없이 진단 결과의 정확성과 일관성을 일정 수준 보장할 수 있다. 본 논문에서는 데이터 분포 통계를 날짜 도메인 데이터로 제한하여 적용하였으나 수량 도메인, 번호 도메인 등의 숫자 형식의 데이터에 적용할 경우 정답률 향상을 기대할 수 있을 것으로 보인다. 추후 연구과제로 데이터 분포 통계 적용 범위 확장에 대한 연구가 진행되어야 할 것이다.

REFERENCES

[1] Yjyoung, International Open Data Status and Open Data Development Direction, Proceedings of Symposium of the Korean Institute of communications and Information Sciences, pp. 529-530, Republic of Korea, June 2015.

[2] Ywhong, "A study on the invigorating strategies for open government data", Journal of the Korean Data And Information Science Society, Vol. 25, No. 4, pp. 769-777, Aug 2014. DOI: http://dx.doi.org/10.7465/jkdi.2014.25.4.769

[3] Kassen, Maxat. "A promising phenomenon of open data: A case study of the Chicago open data project", Government information quarterly, Vol. 30, No. 4, pp. 508-513, Aug 2013. DOI: https://doi.org/10.1016/j.giq.2013.05.012

[4] Hjyi, Yjnam, "A Study on Revitalizing the Use of Korean Public Data: Focused on Linked Open Data Strategy", Journal of the Korean Society for Information Management, Vol. 31, No. 4, pp. 249-266, 2014. DOI: https://accesson.kr/kosim/v.31/4/249/750

[5] Jwlim, Ghchoi, "The Influence of Open Data Policies on Public Innovation", Journal of the Korean Institute of Industrial Engineers, Vol. 43, No. 1, pp. 19-29, Feb 2017. DOI: https://doi.org/10.7232/JKIIE.2017.43.1.01

[6] Eskim, "A Study on the Improvement of the Legal System for the Promotion of Opening and Utilization of Open Government Data - Focusing on cases of refusal to provide -," informatization Policy, Vol. 30, No. 2, pp. 46-67, 2023. DOI:https://doi.org/10.22693/NIAIP.2023.30.2.046

[7] Ministry of Government Legislation, Act on Promotion of Provision and Use of Public Data, https://www.law.go.kr/법령/공공데이터의제공및이용활성화에관한법률/(20201210,17344,20200609)/제2조

[8] Ministry of the Interior and safety, Public data portal, https://www.data.go.kr.

[9] Hlkim, "Quality Evaluation of the Open Standard Data", journal of the Korea Contents Association, Vol. 20, No. 9, pp. 439-447, Sep 2020. DOI: https://doi.org/10.5392/JKCA.2020.20.09.439

[10] Cjkim, gepark, "Quality Characteristics of Public Open Data," Journal of Digital Convergence, Vol. 13, No. 10, pp. 135-146, Oct 2015. DOI: https://doi.org/10.14400/JDC.2015.13.10.135

[11] ShPark, khLee, ayLee. "An Empirical Study on the Effects of Source Data Quality on the Usefulness and Utilization of Big Data Analytics Results" Korea Data Strategy Society, Vol. 24, No. 4, pp. 197-214, Dec 2017. DOI: https://doi.org/10.21219/jitam.2017.24.4.197

[12] Smkim, Jskim, "A Study on the Evaluation Model for Reliability of Public Data," The Journal of Korean Institute of Information Technology, Vol. 21, No. 1, pp. 21-28, 2023. DOI: 10.14801/jkiit.2023.21.1.21

[13] Cesong, Hlkim, "Improvements of public data policy through data portal analysis of local governments," Journal of Digital Contents Society, Vol. 23, No. 4, pp. 697-705, 2022. DOI: 10.9728/dcs.2022.23.4.697

[14] Swkong, Dyhwang, "A Study of Big Data Domain Automatic Classification Using Machine Learning," The Korea Journal of BigData, Vol. 3, No. 2, pp. 11-18, 2018.

[15] Ministry of Government Legislation, Act on Promotion of Provision and Use of Public Data, https://www.law.go.kr/법령/공공데이터의제공및이용활성화에관한법률/(20201210,17344,20200609)/제22조

[16] Ministry of the Interior and Safety, “2022 Public Data Quality Management Level Diagnosis and Evaluation Manual”, pp. 1-59, 2022

[17] Ministry of Government Legislation, Database standardization guidelines for public institutions, https://www.law.go.kr/행정규칙/공공기관의데이터베이스표준화지침/(2023-18,20230403)

[18] Ministry of the Interior and Safety, Information on the 5th enactment of common standard terms for public data, https://zrr.kr/A44p

[19] Ministry of the Interior and safety, “Public data open standards”, 2014.

Authors

Ha-Na Jeong received the B.S., M.S. degrees in Computer Science Engineering from Kongju National University, Cheonan, in 2019, 2021 respectively. She is currently pursuing a Ph.D. degree in Computer Science Engineering from Kongju National University, Cheonan. She is interested in database, data quality, quality diagnosis.

Jae-Woong Kim received the bachelor's degree and the M.S. degree in the Department of Computer Engineering from the Jungang University in 1983 and 1988, respectively. He received the Ph.D. degree in the Department of Computer Engineering from Daejun University in 2000. He has been a professor in the Department of Computer Engineering at Kongju National University since 1992. His current research interests include software engineering.

Young-Suk Chung received the M. S. degree in Multimedia Engineering from Kongju national university, in 2009. Ph. D degree in Computer Engineering from Kongju national university in 2013. He is currently an adjunct professor in Daejeon Health Sciences College. He is interested in Big data, Cloud computing, Simulation, A.I and Predictive modeling.