The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 22, No. 4, pp.81-86, Aug. 31, 2022. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2022.22.4.81
JIIBC 2022-4-12

공공데이터의 도메인 자동 판별 정확도 향상을 위한 정규표현식 및 접미사 적용 방법

Application Method of Regular Expressions and Suffixes to improve the Accuracy of Automatic Domain Identification of Public Data

김석균*, 이관우**

Seok-Kyoun Kim*, Kwanwoo Lee**

요 약 본 연구에서 csv포맷으로 구조화된 파일 데이터의 컬럼의 도메인을 자동 판별하는 방법을 제안한다. 데이터와 데이터 간 융합을 통해 새로운 데이터를 생성할 수 있고, 이들 새로운 데이터가 중요한 자원이 되기 위해서는 조인 되는 컬럼의 일관성이 유지되어야 한다. 데이터 품질을 측정하기 위한 방법 중의 하나가 도메인 기반 품질 진단 방법이다. 도멘인이란 각 컬럼의 성격을 규정하는 가장 광범위한 지표이므로 이를 자동으로 판별하는 방법이 필요하다. 기존의 연구에서는 관계형 데이터베이스의 도메인 자동 판별이 주로 연구 되었지만 본 연구는 파일데이터의 특성을 이용하여 도메인을 자동화 할 수 있는 모델을 개발하였다. 파일데이터의 도메인 판별을 특화하기 위하여 정규표현식을 이용하여 데이터를 단순화 하고 이를 패턴화 하였고, 컬럼명에 해당하는 데이터 헤더의 내용을 분석하여 사용된 접미사를 분석하여 파생변수로 사용하였다. 정규표현식과 접미사의 파생변수를 추가하였을 때 기존 방법인 87%의 정확도 보다 큰 95%의 정확도로 도메인을 자동 판별하는 결과를 도출하였다. 본 연구는 공공데이터 품질진단에 자동화 방법론을 제시하여 품질 측정 기간 및 인원을 줄일 수 있을 것으로 기대된다.

Abstract In this work, we propose a method for automatically determining the domain of columns of file data structured by csv format. New data can be generated through convergence between data and data, and the consistency of the joined columns must be maintained in order for these new data to become an important resource. One of the methods for measuring data quality is a domain-based quality diagnosis method. Domain is the broadest indicator that defines the nature of each column, so a method of automatically determining it is necessary. Although previous studies mainly studied domain automatic discrimination of relational databases, this study developed a model that can automate domains using the characteristics of file data. In order to specialize in the domain discrimination of file data, the data were simplified and patterned using a regular expression, and the contents of the data header corresponding to the column name were analyzed, and the suffix used was used as a derived variable. When derivatives of regular expressions and suffixes were added, the result of automatically determining the domain with an accuracy of 95% greater than the existing method of 87% was derived. This study is expected to reduce the quality measurement period and number of people by presenting an automation methodology to the quality diagnosis of public data.

Key Words : AI, Data Quality, Data Architecture, Regular Expression

*정회원, 한성대학교학교 스마트융합컨설팅학과
**정회원, 한성대학교학교 AI응용학과
접수일자 2022년 6월 11일, 수정완료 2022년 7월 11일
게재확정일자 2022년 8월 5일

Received: 11 June, 2022 / Revised: 11 July, 2022 /
Accepted: 5 August, 2022
*Corresponding Author: kwlee@hansung.ac.kr
Division of IT Convergence Engineerin, Hansung University, Korea.

Ⅰ. 서 론

인공지능의 발전으로 우리는 정보화시대를 넘어 지능정보 시대로 변경되고 있다. 지능정보 시대에는 데이터를 활용해 공공서비스, 경제 및 산업 등 사회 전반에 혁신을 이루기 위해서 국가 차원의 데이터에 대한 전략적 계획이 중요하게 될 것이다. 이에 따라 대한민국 정부는 데이터의 생성부터 활용에 이르기까지 공공데이터의 개방·활용의 중요성과 관련된 정책을 강조하고 있다. 공공데이터란 공공기관이 업무 수행의 결과로 생성 또는 취득한 모든 자료를 말한다.

데이터는 이제 아주 중요한 자원이다. 데이터와 데이터를 합성하여 새로운 데이터를 생성할 수 있고, 이렇게 합성된 데이터가 중요한 자원으로 활용될 수 있다. 데이터와 데이터를 합성하기 위해서는 연결되는 데이터 항목의 형식에 대한 정확성과 일관성이 유지되어야 한다 [1].

도메인 기반 데이터품질 진단방법은 데이터 항목의 형식에 대한 정확성과 일관성을 확보하기 위해서 데이터 항목과 관련된 도메인을 기반으로 항목 데이터가 도메인의 특성에 부합하는지를 판별한다[2]. 도메인 기반 데이터 품질 진단을 수행하기 위해서는 데이터 항목과 관련된 도메인의 판별이 선행되어야 한다.

지금까지 데이터 항목과 관련된 도메인 판별은 수작업으로 진행되어 왔다. 정부는 청년 인턴십과 같은 사업을 통해서 대규모의 공공데이터에 대해 도메인 판별을 통한 품질 진단을 수작업으로 진행해 왔으나 비용 및 시간이 많이 드는 한계가 존재한다. 최근 들어 관계형 데이터베이스 형식의 데이터에 대한 도메인 자동 판별 방법[3, 4]이 제안되었으나, 파일 형식의 공공데이터에 대한 도메인 자동 판별 방법은 부재하다.

본 연구에서는 기존 관계형 데이터베이스 기반의 도메인 자동 판별 방법을 확장하여 csv파일 형식의 공공데이터에 대해서도 적용 가능한 방법을 제안한다. 이를 위해서 2장에서는 본 연구의 이론적 배경인 도메인 기반 데이터 품질진단 및 기존 도메인 자동판별 방법에 대해서 간략히 설명하고, 이전 연구의 한계점 및 개선점에 관해 서술한다. 3장에서는 본 논문에서 제안한 파일 데이터에 대한 도메인 자동 판별 방법에 대해서 구체적으로 기술한다. 4 장에서는 제안한 방법을 적용한 실험결과와 함께 분석 내용을 기술한다.

Ⅱ. 이론적 배경

1. 도메인 기반 데이터 품질진단 방법

데이터 품질이란 “데이터를 활용하는 사용자의 다양한 활용 목적이나 만족도를 지속적으로 충족시킬 수 있는 수준”이라 정의된다. 데이터 품질을 측정하고, 데이터 품질의 신뢰성이 낮은 원인을 파악하고 개선하는 과정을 통해 데이터 이용자의 만족도를 극대화하기 위해 수행하는 일련의 과정이 데이터 품질을 측정하는 이유이다 [1].

데이터 품질은 과거보다 현재에 더 중요도가 커지고 있다. 특히, 방대한 데이터를 활용하는 기계학습이 올바른 결과를 산출하기 위해서는 사용되는 데이터의 품질에 대한 중요도가 더욱 높아지고 있다[7].

데이터 품질 기준은 데이터 컬럼 값에 누락이 없어야 하는 완전성과, 컬럼 값은 유일해서 중복이 없어야 하는 유일성, 컬럼 값이 정해진 데이터 유효 범위 및 도메인을 충족해야 하는 유효성, 데이터가 지켜야 할 구조 및 형태가 일관되게 정의되고 유지되어야 하는 일관성, 그리고, 실세계에 존재하는 값이 정확히 반영되어 있는지에 대한 정확성으로 정의할 수 있다[5].

도메인 기반 데이터 품질진단 방법은 유효성과 일괄성을 관점에서 데이터가 해당 도메인 특성에 부합하는 지를 분석한다.

도메인이란 관계형 데이터베이스에서 테이블에 속한 컬럼들의 고유한 특성을 의미한다[6]. 도메인은 크게 문자, 숫자 등으로 구성되어 일정한 패턴을 갖는 번호도메인, 사물이나 사람의 이름이나 내용을 표시하는 명(내용)도메인, 금액, 비율, 수치, 값 등을 나타내는 숫자로 표기되는 수치도메인, 그리고 날짜나 기간을 포함하는 날짜도메인과 여부나 유무 등 2개 중 한 개를 선택할 수 있는 여부도메인으로 구분된다.

표1는 도메인을 기반으로 데이터 품질진단을 수행했을 때의 결과 예시를 나타낸다. 여부도메인의 경우는 데이터가 「Y」또는 「N」의 형태를 유지하는지를, 번호도메인이나 날짜도메인의 경우에는 정해진 번호나 날짜 포맷의 형태가 일관적으로 유지되는지를 분석한다. 또한, 수치도메인의 경우에는 일반적으로 음수(-)가 포함된 숫자형이지만 「판매가격」과 같은 도메인의 경우 판매가격은 0보다 작은 값을 가지는 것은 의미론적으로 판단이 필요한 경우도 있다.

표 1. 데이터 진단 오류 예시
Table 1 . Examples of data evaluation errors

도메인	컬럼명 예시	컬럼값	오류 여부	오류 원인
여부	고객여부	Y	N	
		N	N	
		1	Y	Y, N 외 문자
		0	Y	
		O	Y	
		X	Y	
번호	전화번호 (하이픈패턴)	02-1234-5678	N	
		02 1234 5678	Y	하이픈 없음
		-	Y	데이터 null
		11111112	Y	전화번호 아님
		2015-11-13	Y	전화번호형식 아님
수치	판매가격	-9999999	Y	판매가격이 0보다 작을수 없음
		오백만원	Y	한글로 입력
		1000000	N	
날짜	배송일자 (YYYYMMDD)	20220403	N	
		201506	Y	YYYYMM
		2015	Y	포맷이 YYYY
		20220231	Y	2월은 31 없음
		2015-11-13	Y	하이픈 추가

2. 기계학습을 통한 도메인 자동 판별 연구

도메인 기반 데이터 품질진단 방법을 적용하기 위해서는 진단할 데이터 컬럼과 관련된 도메인의 판별이 선행되어야 한다. 도메인 자동 판별을 위한 기존 연구[4]는 의사결정 트리 알고리즘을 변형한 랜덤포레스트 알고리즘을 사용하는 기계학습 방법을 바탕으로, 관계형 데이터베이스와 SQL을 이용하여 추출할 수 있는 정보만을 파생변수로 선택하였다. 표2와 같이 이전연구의 파생변수들은 관계형 데이터베이스의 데이터타입 및 데이터의 소수점 길이, 날짜 형식의 데이터타입 등의 사용 여부 등이 중요한 파생변수로 사용되었다.

공공데이터의 경우 파일 형태로 존재하는 경우가 대부분이고 또한 표 형식이기 때문에 관계형 데이터베이스의 테이블 형식으로 변환이 비교적 수월하다. 따라서 기존 도메인 자동 판별 방법[4]을 파일 형식의 공공데이터에 적용하여 진행할 수도 있다.

하지만, 기존 방법은 파생변수를 관계형 데이터베이스 기반에서 적합한 항목만을 선택하였기 때문에 파일 형태로 존재하는 공공데이터에서 일부 파생변수는 데이터값을 가지지 못하는 경우도 존재한다. 예를 들면, 「데이터타입」파생변수는 데이터타입이 정의되지 않는 파일 데이터에서는 적용되기 힘들고, 「PK여부」파생변수도 데이터베이스의 테이블 구조와 관련된 항목이기 때문에 적용이 힘들다.

또한, 「날짜형식여부」파생변수는 데이터가 날짜 형식인지 아닌지 만을 판별할 수 있으므로, 공공데이터에서 나타나는 다양한 형태의 날짜형식(예, 연월일, 연월일시분초)으로 표현되는 데이터의 도메인을 세분화하여 판별하는 데는 미흡하다.

다음 장에서는 기존 기계학습 기반의 도메인 자동 판별 연구가 파일 형태의 공공데이터에 바로 적용되기에 어려운 한계점을 극복하기 위한 방법을 제안한다.

표 2. 관계형 데이터베이스 데이터를 위한 파생변수
Table 2. Dervied variables for relational database data

파생변수	설명
데이터타입	INT, CHAR, VARCHAR 등 같은 데이터 값을 구분할 수 있는 변수
데이터 최대길이	칼럼 내의 데이터 중 최대 길이를 가지고 있는 데이터의 길이
데이터길이변화	칼럼 내의 데이터 길이의 가변 여부
소수점아래길이	칼럼 내의 데이터들의 소수점 아래의 길이 데이터
날짜형식여부	데이터 타입이 아닌 날짜 포맷 데이터 여부
연락처형식여부	@, - 등 연락처 및 주소에서 사용하는 패턴을 이용한 데이터 존재여부
공백비율	전체 데이터에서 공백이 차지하는 여부
엔터포함여부	칼럼 내의 데이터에서 줄바꿈이 일어났는 지 여부
영어작성여부	데이터들이 영어로만 작성되었는지 여부
숫자작성여부	데이터들이 숫자로만 작성되었는지 여부
백단뒤이하비율	칼럼 내의 데이터 중 100단위 이하는 000으로 표기 된 비율
그룹화비율	칼럼 내의 데이터 중 그룹화가 가능한 비율
PK여부	칼럼이 Primary Key로 설정되었는 지 여부

Ⅲ.파일 데이터의 도메인 자동 판별

파일 데이터의 경우 도메인을 판별할 수 있는 근거가 맨 첫 번째 행의 컬럼 명에 의해 해당 컬럼의 특징을 유추할 수 있다. 또한, 관계형 데이터베이스와는 달리 데이터타입이 없으므로, 데이터를 최소한 읽어서 그 형태가 어떻게 유지되거나 변하는지 파악하는 프로파일링이 필요하다. 예를 들어, 날짜도메인의 경우에 「22-03-02」,

「2022-03-22」,「2022년3월22일」,「2022-03-22 14:20:12」와 같이 같은 날짜임에도 여러 가지 형식이 달라질 수 있으므로, 이런 날짜도메인을 판별하는 것이 세분화될 필요가 있다.

따라서 본 논문에서는 기존 기계학습 기반의 도메인 자동 판별 연구를 파일 형식에 공공데이터에 확장 적용할 수 있도록 새로운 파생변수를 추가한다.

1. 컬럼명 접미사의 파생변수

컬럼명에 해당하는 첫 행은 주로 명사형의 형태를 띠고 있으므로 한국어의 특성상 접미사를 통해 컬럼의 형식을 유추할 수 있다. 예를 들어「~금액」 등으로 끝나는 접미사는 수치도메인의 가능성이 크며「~일자, ~시간」은 날짜도메인 그리고 「~여부」는 여부 도메인일 가능성이 크다. 따라서, 도메인별 관련 접미사를 표 3과 같이 정의하고, 이를 도메인을 판별하는 파생변수로 사용한다.

표 3. 도메인 별 접미사 예시
Table 3. Examples of suffixes by domain

도메인	접미사
여부	여부, 유무
명	명, 내역, 내용, 이름, 구분, 유형
전화번호	전화, 전화번호, 문의및안내, 연락처, 팩스, FAX, FAX번호, 해피콜번호, 고객센터, 문의처, 휴대폰, 휴대전화, 휴대폰번호, 이동통신번호
우편번호	우편번호, ZIP, 우편, ZIPCODE, ZIPCD
사업자번호	사업자번호, 사업자등록번호, 사업자번
날짜	연월, 년도, 연도, 일시, 년, 일자, 연월중, 시간, 연, 년월일, 기간, 월별, 월, 기준일, 시기, 날짜, 일, 시각, 초, 시, 분
수치	금액, 가격, 잔액, 비용, 금리, 비율, 이자, 이율, 수량, 소계, 인원, 면적, 건수, 페이지, 합계, 평균, 최소값, 최대값, 표준편차, 율, 률, 수, 량, 건, 개, 비, 가, 원, 액, 료, 금

2. 정규표현식 패턴의 파생변수

본 연구에서는 정규표현식을 이용한 파일 데이터의 패턴을 정의하고 이를 바탕으로 도메인을 판별하는 방법을 확장한다.

정규 표현식은 특정 규칙을 가진 문자열을 가장 효과적으로 표현할 수 있는 언어로, 애플리케이션과 프로그래밍 언어에 사용할 수 있는 특수한 텍스트 패턴이다.

전화번호의 경우 「02-123-4567」,「02-1234-5678」, 「010-123-5678」,「010-1234-5678」는 다음과 같은 정규 표현식으로 구성된다.

```
^[0-9]{2,3}-[0-9]{3,4}-[0-9]{4}$        (1)
```

또한, 숫자는 모두 '9', 소문자 영문은 'a', 대문자는 'A',한글은 'H'로 표현하고 기호는 그대로 사용을 하면 정규 표현식 식1은 다음 (2)와 같이 더 단순화된다.

```
^9{2,3}-9{3,4}-9{4}$        (2)
```

표의 형식이 대부분인 공공데이터의 경우 한 컬럼에 대해 여부, 날짜, 전화번호, 수치 도메인을 포함하는 데이터는 특정 패턴이 존재한다.

명 도메인이나 내용도메인이 포함된 문자열 데이터는 길이 등에 많은 변화가 있기 때문에 패턴에서 제외한다.

표 4는 공공데이터에서 가장 많이 사용되는 도메인에 대해서 정규표현식의 형태로 정의한 것을 나타낸다.

표 4. 도메인 별 정규표현식 예시
Table 4. Examples of regular expressions by domain

도메인	정규표현식
여부	A
전화번호	^(\-)*9{1,15}(\.)*(9{1,15})*%*$
우편번호	^999-999$
사업자번호	^(9999999999|999-99-99999)$
날짜 : 년	^9999H{0,1}$
날짜 : 년_월	^9999[H|-|:|-]{0,1}9{1,2}H*$
날짜 : 월-일	^99[H|-|:|-]9{1,2}H*$
날짜 : 년-월-일	^9999[H|-]9{1,2}[H|-]9{1,2}$
날짜 : 시:분:초	^99[H|:]9{1,2}[H|:]9{1,2}$
날짜 : 년-월-일 시:분	^9999-9{1,2}-9{1,2}B99:99$
날짜 : 년-월-일 시:분:초	^9999-99-99B99:99:99(\.9{1,7})*$
수치	^(\-)*9{1,15}(\.)*(9{1,15})*%*$

컬럼의 데이터값을 프로파일링하여 표4의 정규표현식에 적용되는 데이터의 비중이 50%가 초과한 경우 이를 해당 패턴으로 처리하여 도메인 판별에 사용한다.

3. 데이터 수집 및 도메인 판별 프로세스

데이터는 공공데이터 포털에 공개된 파일 데이터 중 다운로드를 하여 온전한 csv 포맷을 유지하고 있는 8440개의 파일을 다운로드 하여 93,286개의 컬럼을 최대 500라인의 데이터를 프로파일링 하여 데이터를 추출했다.

전처리가 불가능한 문자나 컬럼명이 공백으로 처리된 불량한 데이터는 제외를 하여 총 13만개의 컬럼 내용중 데이터 처리가 가능한 93,286개의 컬럼을 대상으로 데이터를 수집했다.

훈련 데이터와 검증데이터는 8:2의 비율로 분리해서

사용했고 기계학습을 위해 파일당 500라인을 프로파일링한 결과를 데이터로 이용하였다.

분류모델은 의사결정트리 알고리즘과 랜덤포레스트 알고리즘을 이용하여 기존연구의 파생변수와 접미사와 정규표현식의 패턴을 파생변수로 추가했을 때의 상관관계를 비교하였다.

Ⅳ. 연구 결과

기존 연구 파생변수만을 사용하여 도메인 판별을 한 경우 기존 연구의 결과와 비슷한 87%의 정확도를 얻었다. 이는 기존 연구가 관계형 데이터베이스뿐만 아니라 공공데이터 파일 데이터의 도메인 판별에도 적용이 된다는 것을 알려준다. 한편 기존방법에 정규표현식 패턴을 적용하였을 때 95%의 정확도로 약 7%이상 의미있는 정확도의 개선이 보여진다. 또한 세분화된 날짜 도메인도 정확도가 향상되었다.

표5는 랜덤포레스트를 이용한 결과를 기존 방법과 패턴을 적용했을 때의 결과를 비교하였다. 평균제곱오차는 실제값과 예측값의 차이를 기준으로 오차를 판단하는 방식으로 값이 작을수록 오차가 줄어드는 것을 나타낸다.

$$\text{평균제곱오차} = \frac{1}{n}\sum(y-\hat{y})^2 \qquad (3)$$

표 5. 정규표현식 패턴 적용전과 적용 후 표준편차 비교
Table 5. Comparison of standard deviation before and after applying regular expression patterns

	기존방식	패턴적용
평균제곱오차	4.77	2.14
설명분산점수	0.62	0.83
정확도	0.87	0.95

설명분산점수는 예측된 결과의 분산 비율을 측정하는데 사용되는데 측정값과 예측값의 1에 가까울수록 예측이 정확해 진다. 표5와 같이 패턴을 적용했을 때 평균제곱오차가 줄어들고, 성명분산점수가 향상되어 정확도가 증가했음을 알 수 있다.

패턴을 적용하지 않았을 때와 패턴을 적용한 경우의 파생변수의 중요도를 그림1과 2에 표시하였다.

패턴 미적용 시 중요 요인은 데이터의 최대 길이와 최소길이가 중요하지만, 패턴 적용의 경우 정규표현식을 이용한 패턴과 접미사의 패턴을 적용이 두드러지게 중요한 요인이 되었다.

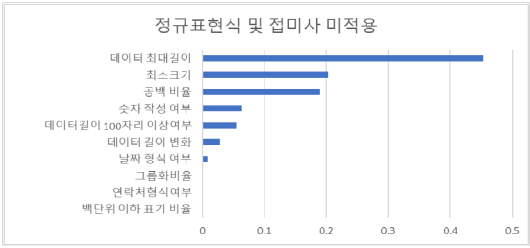

그림 1. 정규표현식 및 접미사 미적용시 파생변수 중요도
Fig. 1. Importance of derivatives when regular expressions and suffixes are not applied

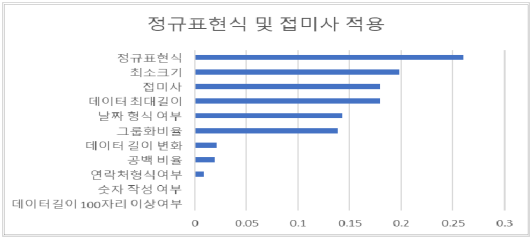

그림 2. 정규표현식 및 접미사 적용시 파생변수 중요도
Fig. 2. Importance of Derivative Variables in the Application of Regular Expressions and Suffixes

표 6. 도메인별 정확도 (정규표현패턴 미적용시 vs. 적용시)
Table 6. Accurarchy for each domain (when regular expression pattern is not applied vs. when applied)

	미적용			적용		
	precision	recall	f1-score	precision	recall	f1-score
문자열	0.83	0.84	0.84	0.94	0.93	0.93
수치	0.91	0.91	0.91	0.96	0.97	0.96
여부	0.4	0.02	0.03	0.96	0.84	0.9
전화번호	0.87	0.88	0.87	0.97	0.94	0.96
우편번호	0	0	0	0.59	0.5	0.54
사업자번호	0	0	0	1	0.67	0.8
날짜_HH24	0.63	0.44	0.52	0.85	0.6	0.7
날짜_HH24:MI	0.98	1	0.99	1	1	1
날짜_MM-DD HH24:MI	0.49	0.92	0.64	0.87	0.92	0.89
날짜_YYYY	1	0.84	0.91	1	0.95	0.97
날짜_YYYY_MM	0.94	0.97	0.96	1	0.98	0.99
날짜_YYYY-MM-DD	0.96	0.24	0.38	1	0.99	0.99
날짜_YYYY-MM-DD HH24:MI	0.69	0.79	0.74	0.98	1	0.99

그룹화 비율이 늘어난 이유는 기존 데이터베이스의 형태에서는 특정 패턴을 적용하여 그룹화하지 않고 데이터의 크기나 타입 등으로 분류가 되었고, 정규표현식 패턴이 적용되면 패턴별로 그룹화가 이루어지기 때문에 그룹화 비율의 비중이 높아지고 이 파생변수가 전과는 다르게 의미가 있게 됨을 알 수 있다.

V. 결 론

데이터의 활용도가 높아지고 데이터의 합성은 4차혁명시대에 반드시 필요한 새로운 미래형 먹거리가 될 것임에 자명하다. 데이터는 매 순간 우리가 알지도 못하는 사이에 엄청나게 쌓이게 된다. 공공데이터 뿐 아니라 빅데이터를 보관하는 가장 좋은 방법은 파일형태로 처리하는 것이다. 데이터와 데이터간 매시업을 통해 새로운 데이터를 만들기 위해 데이터와 데이터 사이의 연결 매개가 될 수 있는 도메인을 판단하고 그 도메인에 대한 형식을 통일 시킨다면 데이터 매시업은 보다 더 수월하게 처리 될 수 있다.

본 논문에서 파일데이터 도메인 판별을 위해 정규표현식을 이용한 패턴을 이용한 기계학습 방법을 제안하였고, 파일형태의 공공데이터 도메인의 판별을 위해 파생변수에 정규표현식의 패턴을 추가하여 결과를 확인 한 결과 기존의 데이터베이스적인 관점의 파생변수 방법보다 약 7%정도의 정확도가 향상이 되어 유의미한 결과를 도출하였다. 표6과 같이 전체적으로 모든 도메인에대해 유의미한 정확도가 증가되었다.

향후 파일데이터의 컬럼간 유사도를 측정하고 수치화하여 정규표현식의 패턴과 적용을 한다면 보다 더 유의미한 데이터 품질 진단을 예측한다.

References

[1] Ko, Kwangman, and Park, Hong-Jin. “Development of the Pattern Matching Engine Using Regular Expression.” The Journal of the Korea Contents Association, Vol. 8, No. 2, pp 33-40, 2008.
DOI: https://doi.org/10.5392/JKCA.2008.8.2.03

[2] Chae, Cheol Joo, and Hong, Eui Kyeong. "Quality Management Model of Atypical Science and Technology Big Data Based on Data Profiling and Regular Expression" The Journal of the Korea Contents Association, Vol. 14, No.12 ,pp. 486-493, 2014.
DOI: https://doi.org/10.5392/JKCA.2014.14.12.486

[3] Jin-Hyoung Lee, "A Study on Automation of Big Data Quality Diagnosis Using Machine Learning", The Journal of Bigdata, Vol.2, No.2, pp.75 - 86, 2017.
DOI: https://doi.org/10.36498/kbigdt.2017.2.2.75

[4] Kong Seongwon, Hwang Deokyoul, "A Study of Big Data Domain Automatic Classification Using Machine Learning", The Journal of Bigdata,, Vol.3, No.2, pp. 11-18, 2018.
DOI: https://doi.org/10.36498/kbigdt.2018.3.2.11

[5] Korea Database Agency's editorial department, "Data Quality Guidelines", Korea Database Agency, August 10, 2011

[6] Hong-Jin Park, “Trend Analysis of Korea Papers in the Fields of ‘Artificial Intelligence’, ‘Machine Learning’ and ‘Deep Learning’”, Journal of Korea Institute of Information,Electronics,and Commu- nication Technology, Vol. 13, No. 4, pp. 283-292, 2020.
DOI: https://doi.org/10.17661/jkiiect.2020.13.4.283

[7] Cha, Kyung-Yup, and Sim, Kwang-Ho. “A Methodological Framework for Assessing the Reliability of Computer-Processed Data.” Commu- nications for Statistical Applications and Methods, Vol. 17, No. 5, pp. 745-753, 2010.
DOI: https://doi.org/10.5351/CKSS.2010.17.5.745.

저 자 소 개

김 석 균(정회원)

- 1997년 : 수원대학교 고분자공학과 (학사)
- 2000년 : 수원대학교 고분자공학과 (공학석사)
- 2008 ~ 현재 : ㈜위세아이텍 수석
- 2020 ~ 현재 : 한성대학교 스마트융합학과 박사과정

- 관심분야 : 데이터아키텍처,소스트웨어프로덕트라인

이 관 우(정회원)

- 1994년 : 포항공과대학교 전자계산학과(학사)
- 1996년 : 포항공과대학교 컴퓨터공학과(공학석사)
- 2003년 : 포항공과대학교 컴퓨터공학과(공학박사)
- 2003년 ~ 현재 : 한성대학교 AI응용학과 교수

- 관심분야 : 소프트웨어 제품계열 (Software Product Line), 관점지향 프로그래밍(Aspect-Oriented Programming), 소프트웨어 아키텍처

※ 본 연구는 한성대학교 학술연구비 지원과제 임