JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 26 No. 4, pp. 105-112, April 2021
https://doi.org/10.9708/jksci.2021.26.04.105

A Study on the Classification of Unstructured Data through Morpheme Analysis

SungJin Kim*, NakJin Choi*, JunDong Lee*

*Ph.D. ing, Dept. of Multimedia Engineering, GangNeung-Wonju National University, Wonju, Korea
*Ph.D. ing, Dept. of Multimedia Engineering, GangNeung-Wonju National University, Wonju, Korea
*Professor, Dept. of Multimedia Engineering, GangNeung-Wonju National University, Wonju, Korea

[Abstract]

In the era of big data, interest in data is exploding. In particular, the development of the Internet and social media has led to the creation of new data, enabling the realization of the era of big data and artificial intelligence and opening a new chapter in convergence technology. Also, in the past, there are many demands for analysis of data that could not be handled by programs.

In this paper, an analysis model was designed and verified for classification of unstructured data, which is often required in the era of big data. Data crawled DBPia's thesis summary, main words, and sub-keyword, and created a database using KoNLP's data dictionary, and tokenized words through morpheme analysis. In addition, nouns were extracted using KAIST's 9 part-of-speech classification system, TF-IDF values were generated, and an analysis dataset was created by combining training data and Y values. Finally, The adequacy of classification was measured by applying three analysis algorithms(random forest, SVM, decision tree) to the generated analysis dataset.

The classification model technique proposed in this paper can be usefully used in various fields such as civil complaint classification analysis and text-related analysis in addition to thesis classification.

▸ **Key words**: Big Data, Data Analysis, Visualization, Textmining, Modeling

[요 약]

빅데이터 시대에 접어들며 데이터에 대한 관심이 폭발적으로 늘어나고 있다. 특히, 인터넷 및 소셜미디어의 발전은 새로운 데이터들의 생성으로 연결되어 빅데이터와 인공지능 시대의 실현과 융합 기술의 새로운 장을 열 수 있게 되었으며, 과거에는 프로그램으로 다루지 못하던 데이터에 대한 분석 요구가 많이 발생하고 있다.

본 논문에서는 빅데이터 시대에서 많이 요구되는 비정형 데이터에 대한 분류를 위하여 분석 모델을 설계하고 이를 검증하였다. 데이터는 디비피아의 논문 요약과 주제어, 그리고 부주제 어를 크롤링하였으며, 코엔엘피의 데이터 사전을 이용해 데이터베이스를 생성하고, 형태소 분석을 통하여 단어의 토큰화 과정을 수행하였다. 또한, 카이스트의 9 품사 분류 체계를 이용해 명사를 추출하고, TF-IDF 값을 생성하였으며, 학습 데이터와 Y 값을 결합하여 분석 데이터 셋을 생성하였다. 이와 같이 생성된 분석 데이터 셋에 랜덤 포레스트와 서포트 벡터 머신 그리고 의사결정트리, 이렇게 세 가지 분석 알고리즘을 적용하여 분류의 적정성을 측정하였다.

본 논문에서 제안한 분류 모델 기법은 논문 분류 외에도 민원 분류 분석 및 텍스트 관련 분석 등 다양한 분야에 유용하게 사용될 수 있다.

▸ **주제어**: 빅데이터, 데이터분석, 시각화, 텍스트마이닝, 모델링

• First Author: SungJin Kim, Corresponding Author: JunDong Lee
*SungJin Kim (tonyksj@naver.com), Dept. of Multimedia Engineering, GangNeung-Wonju National University
*NakJin Choi (nakjin@ex.co.kr), Dept. of Multimedia Engineering, GangNeung-Wonju National University
*JunDong Lee (jlee@gwnu.ac.kr), Dept. of Multimedia Engineering, GangNeung-Wonju National University
• Received: 2021. 03. 04, Revised: 2021. 03. 29, Accepted: 2021. 03. 29.
• This work is the extended research of 2020 KSCI Summer Conference(p253-255)

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

I. Introduction

빅데이터 시대에 접어들며 데이터에 대한 관심이 폭발적으로 늘어나고 있다. 이런 빅데이터 시대에서 데이터는 정형 데이터에 비해 비정형 데이터의 규모가 방대하고, 생성 주기 또한 짧아지고 있는 추세이다.

과거 빅데이터는 데이터로서의 가치보다는 관리의 측면에서 활용되었다. 한 예로 보면 서버의 로그 데이터는 시스템 이상 시 문제를 해결하는 용도였기에 용량의 한계에 도달하면 삭제를 하곤 했었다. 또 다른 예로는 의료용 기기들의 경우, 발생한 데이터들을 단지 판독용으로만 사용하고 그냥 버려지는 경우가 다반사였다. 하지만 빅데이터 시대가 도래하면서 이제 “버려지는 데이터는 최소한 없다”라는 것이 지론이 되었다. 여기에 각종 소셜 네트워크상에서 발생하는 데이터와 동영상 및 지리정보, 센싱 데이터 등 세상에는 어마어마한 양의 데이터들이 현시점에도 무서운 속도로 발생하고 있다. 이렇듯 엄청나게 늘어나는 데이터를 정확하게 수집하고, 분석, 활용하는 것은 현재까지도 매우 어려운 일이다. 하지만 빅데이터 시대에는 시스템 아키텍처나 분석 기술 등 데이터를 수집 하거나 가공 및 분석하는 기술이 날로 발전하며 거대한 데이터를 통해 가치 분석을 하기 시작하였다.

최근 발생하고 있는 빅데이터의 규모를 보면 1분마다 구글 변역기는 69,500,000 단어를 번역하고 있고, 인터넷 영화관인 넷플릭스(Netflix) 같은 경우에는 86,805 시간의 비디오가 생성되고 있다. 또한, 유튜브 같은 경우에는 100 시간의 새로운 비디오가 생성되고, 드롭박스(Dropbox)에서는 833,333개의 새로운 파일 들이 생성되고 있다[12].

Fig. 1.은 빅데이터의 특징 변화 과정을 분류한 것으로 초기 발표된 빅데이터의 특징으로는 3V(Volume, Velocity, Variety)[09]였으나 최근에는 6V(Volume, Velocity, Variety, Veracity, Value, Visualization) [10]로 변화되며 확대되고 있다. 이렇듯 빅데이터는 시대의 변화에 빠르게 대응하며 변화하고 있는 것이다. 그중에서 데이터의 양(Volume)에 해당하는 데이터는 과거 수치, 즉 코드성 위주의 데이터에서 문자와 영상, 로그, 센싱 데이터 등 다양한 형태의 데이터로 확대되고 있으며, 그 양 또한 과거 발생 데이터의 수 배에서 수천 배 규모로 방대해 지고 있다.

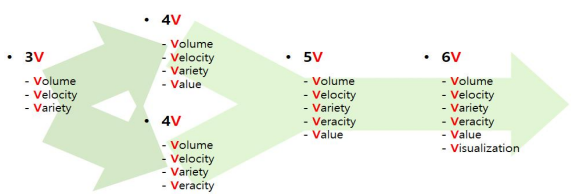

Fig. 1. Big Data Attribute

분석 관점에서도 과거 수치 위주 분석에서 벗어나 시대적 변화에 따른 소셜 네트워크상의 텍스트 등을 분석함으로써 사용자의 생각이나, 패턴 등을 파악하고 대중과 같은 불특정 다수의 흐름과 변화에 대응하는 높은 수준의 분석으로 변화하고 있다.

이렇게 급변하는 시대에 분석의 질을 높이기 위하여 방대한 데이터 중 텍스트 데이터에 대한 분석의 요구가 늘어나고 있는 실정이다.

본 논문에서는 비정형 데이터에 대하여 형태소 분석과 상관관계 분석을 수행하여 자동으로 데이터를 분류하는 시스템을 설계 및 모델링 하였다.

또한, 시스템을 시험하기 위하여 디비피아(DBpia)에서 제공하는 각 논문의 요약을 크롤링을 이용해 6,002건의 데이터를 수집한 후, 전처리 작업을 통해 중복 등을 제거한 3,628건의 논문 데이터를 활용하여 논문의 분류 작업의 적정성을 판단한다.

이를 위하여 2장에서는 비정형 데이터의 처리 방법에 관하여 설명하였고, 3장에서는 본 논문에서 제안한 텍스트 마이닝 분석 기법에 관하여 논하였다. 4장에서는 3장의 제안에 관한 테스트를 수행하여 5장에서 분류작업의 적정성을 결론지었다.

향후 본 논문 내용의 기술은 민원 분석 및 텍스트 관련 분석 등에 활용 범위를 넓혀 나갈 수 있다.

II. Preliminaries

본 논문에서 처리하는 데이터는 기존의 정형 데이터와는 속성이 다르므로 데이터를 수집·저장·처리·분석·시각화하는 방법 또한 기존 방법과는 다른 새로운 시도와 응용이 필요하다.

Fig. 2.는 빅데이터를 처리하는 과정을 크게 데이터의 생성·수집·저장·처리·분석·시각화 과정으로 분류한 것이다.

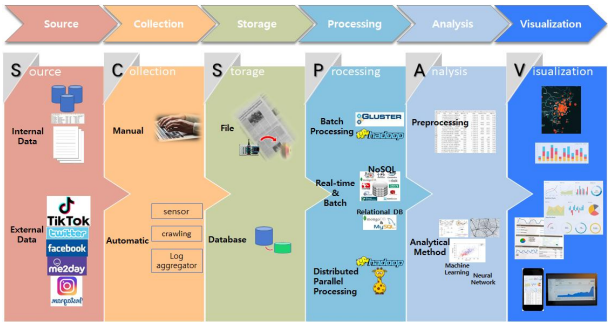

Fig. 2. Big Data Processing

2.1 Source

내부 데이터의 경우는 정형 데이터이므로 인식 및 판단이 빠르고, 결정하기 쉽다. 하지만 외부 데이터의 경우에는 인식과 대상 결정에 상당한 어려움이 있으며 수집하는 과정 또한, 애로사항이 존재하고 있다.

2.2 Collection

데이터를 수집하는 과정은 사람의 개입 여부에 따라서 Table.1과 같이 자동과 수동으로 나눌 수 있다. 데이터 수집 과정은 분석 단계에서 사용되는 데이터를 얻는 중요한 단계이다. 자동의 경우, 전기/전자적인 장치를 이용하며 구축 이후의 비용이 크지 않다는 장점이 있으나 초기 구축이 어렵다는 단점이 있다. 자동으로 수집하는 경우를 보면 핏비트(Fitbit)나 센서에서 발생하는 센싱 데이터 등이 이에 해당하며 사람의 개입 없이 데이터 수집이 이루어지는 경우를 말한다. 수동의 경우에는 사람의 개입이 발생하며 경우에 따라 다양한 데이터 수집이 가능하다는 장점이 있으나 데이터 수집을 위해서는 많은 비용이 발생한다는 단점이 있다. 또한, 수동의 경우는 꼭 사람의 개입이나 노력이 필요하며, 직접 입력하여 데이터를 수집 하거나, 데이터 추출 도구(ETL : Extract Transformation & Loading)를 사용하여 데이터를 수집, 정제하는 방법이 있다[20].

Table 1. Type of data collection methods

Type	Characteristic
Automatic	When data collection takes place without human intervention -Use of mechanical/electronic devices -No additional cost after initial construction.
manual	When human effort must be involved in data collection -When a person manipulates a given device to obtain data -When subjective data is collected through surveys, etc. -The amount of data is proportional to the effort a person puts into collecting necessary data.

2.3 Storage

저장 방법으로는 크게 파일 시스템과 데이터베이스 시스템 방식으로 정의할 수 있다. 파일 시스템의 경우, 빅데이터 시스템의 대표적인 저장 방식으로 HDFS (Hadoop Distributed File System)와 GFS(Google File System) 등이 있으며. 저사양의 서버들을 이용하여 대용량 및 분산처리 등 사용자들에게 고성능의 환경 및 확장성(Scale-out)을 제공한다.

데이터베이스 시스템 방식에는 전통적인 관계형 데이터베이스 방식과 비관계형(NoSQL) 데이터베이스 방식이 있다. 최근 빅데이터 시스템에는 과거 관계형 데이터베이스보다는 비관계형 기반의 데이터베이스를 선호하고 있다.

2.4 Processing

빅데이터 처리 기술에는 일괄처리(Batch Processing) 방식과 실시간처리(Realtime Processing)방식, 그리고 분산병렬처리(Distributed Parellel Processing)방식이 있다. 먼저 일괄처리 방식은 일정 기간, 즉 일이나 월, 연과 같은 단위로 자료를 모아 처리하는 것을 의미한다. 일정 기간 데이터를 모아 두었다가 처리하는 방식이다. 이에 반해 실시간처리 방식은 데이터의 발생 즉시 처리하는 방식이며 온라인상의 스트리밍(Streaming) 데이터 처리 기술을 의미한다. 분산병렬처리 방식은 한 번에 처리해야 할 데이터를 여러 개로 분산하여 처리하는 병렬처리 기술이며, 응답시간을 실시간 수준으로 높이는 방식이다.

2.5 Analysis

예측 분석은 회귀분석, 의사결정트리(Decision Tree)[13], 랜덤 포레스트(Random Forest)[04], 주성분 분석[01][07], 시계열 분석 등과 같이 다양한 기법들이 있다. 회귀분석[05]은 쉽게 분석하고 예측할 수 있다는 장점이 있는 반면, 분석 방법의 선택이나 판단은 분석가에게 달려있어 오용될 수 있다. 의사결정트리는 설명 변수의 규칙이나 관계, 패턴 등을 이용해 목표변수를 분류하는 트리구조의 모델을 생성하여 목표 변숫값의 평균을 예측하는 기법으로, 직관적으로 보기 편하다는 장점이 있다. 랜덤 포레스트는 독립적으로 다수의 모델을 생성하여 결합하고 예측값의 평균을 산출하여 최종모델을 생성하며 큰 데이터에 적용이 가능하며 성능이 뛰어나다는 장점이 있다. 주성분 분석은 변수 간의 관계를 기반으로 정보 손실을 최소화하여 차원을 축소하는 기법으로 다른 분석 기법의 사전 단계로 활용되고 있다. 시계열 분석은 시간의 흐름에 따라 수집된 데이터를 수학적 모델을 이용하여 미래의 값을 예측하는 기법이다.

2.6. Visualization

마지막으로 분석한 결과를 쉽게 이해할 수 있도록 시각적인 수단으로 결과를 전달하는 과정으로서 다양한 시각화를 통해 표출시켜 주는 단계이다. 그래프와 차트, 사진이나 그림 등과 같이 직관적으로 파악이 가능한 요소들로 이루어진다.

III. The Proposed Scheme

본 논문에서는 데이터를 분석하는 방법으로 텍스트 마이닝 기법을 활용하였고, 수행 및 알고리즘 개발을 위하여 Fig. 3.과 같은 과정을 수행하였다.

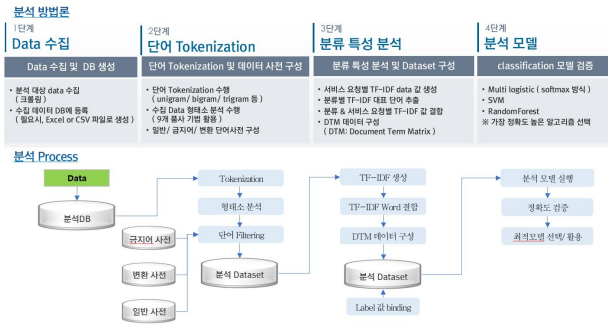

Fig. 3. Textmining Processing

3.1 Data Collection

3.1.1. Definition of Data

데이터를 사전적으로 정의하면 너무 포괄적이므로 좀 더 현실적으로 축소하여 정의해 보면 '모든 분석 또는 활용 가능한 디지털화된 자료, 또는 정보'라고 정의할 수 있다. 데이터가 가지고 있는 특징을 명확하게 인지할 때, 데이터를 수집하고 처리할 수 있는 기술적 고려사항과 처리 과정 설계가 가능하다.

빅데이터의 수집과정은 기존 시스템처럼 데이터를 생산하는 과정보다는 생산된 데이터를 가져오는 수집과정에 비중이 높을 수밖에 없다. 단지 어떤 데이터를 가져올 것인지 보다 생산된 데이터를 수집하는 과정의 안정성이 더 중요하기 때문이다.

수집하는 방법에 따라 자동과 수동으로 나눌 수 있으며, 데이터의 형태에 따라서는 Table. 2.와 같이 정형(Structured data), 반정형(Semi-Structured data), 그리고 비정형(Unstructured data)으로 나눌 수 있다.

Table 2. Comparison of collection difficulty by data type

Type	Characteristic	level of difficulty
Structured data	Most of them are internal systems, so collection is easy. Even a spreadsheet in the form of a file is easy to process	bottom
Semi-structured data	Data processing technology is required because it is usually provided in the form of API.	middle
Unstructured data	In the case of text mining or a file, it is difficult to process the collected data because the file must be parsed in the form of data.	top

또한, Table. 3.과 같이 정형 데이터는 관계형 데이터베이스 시스템과 같이 고정된 컬럼에 저장되는 데이터와 파일, 그리고 지정된 행과 열에 의해 데이터의 속성이 구별되며 데이터의 스키마를 지원한다.

반정형 데이터는 데이터 내부에 스키마에 해당하는 메타데이터를 가지고 있으며, 일반적으로는 파일 형태로 저장된다. 그러므로 반정형의 데이터는 어떤 형태를 가진 데이터인지를 파악하는 것이 필요하다.

비정형 데이터는 데이터 세트가 아닌 하나의 데이터가 수집 데이터로 객체화되어 있으며, 텍스트 데이터나 이미지, 동영상과 같은 멀티미디어 데이터 등이 해당 된다.

Table 3. Comparing the difficulty of architecture configuration by data type

Type	Characteristic	level of difficulty
Structured data	It consists of a general architectural structure where CRUD occurs.	bottom
Semi-structured data	An architectural structure that can be converted into a structured data form by interpreting the meta structure of data must be modified.	middle
Unstructured data	The architecture structure needs to be modified so that text or files can be parsed, converted into a set of meta-structured data, and made into a structured data format.	top

데이터의 유형 파악은 Table. 4.와 Table. 5.과 같이 구성과 존재하는 형태에 따라 데이터 수집 기술을 결정하는 중요한 과정이다[21].

Table 4. Data types by configuration type

Type	Characteristic	Form of existence
Structured data	Data schema supported	RDB, File
Semi-structured data	Meta properties exist in the data	File
Unstructured data	Exist as an image or video file, not a text file and data format that can be analyzed	RDB, File

Table 5. Types of data by type of existence

Type	Data type	Collection method
RDB	Structured data, Unstructured data	DB to DB, ETL, RDB vendor supplied drive
File	Semi-structured data	Crawling, Open API, FTP, HTTP

3.1.2 Collection Techniques

데이터 수집 기술은 데이터 제공 서비스의 품질을 결정하는 중요한 부분이며, 수많은 데이터가 존재해도 수집하는 기술이 미비하다면 서비스에 대한 품질은 기대하기 어려울 것이다.

Table. 6.은 데이터를 수집하는 기술의 종류를 분류한 것으로 크롤링, 오픈 API(Open Application Interface), FTP, RSS(Really Simple Syndication), Streaming, Log Aggregator, RDB Aggregator 등이 있다.

Table 6. Types of technologies in data collection

Type	Characteristic
Crawling	Collects information of web documents provided on the Internet, such as SNS, news, and web information
Open API	Open Application Programming Interface Web programming in the open
FTP	Sending/receiving various files from Internet server using TCP/IP protocol
RSS	Really Simple Syndication Content presentation method commonly used in news or blog sites
Streaming	Real-time collection of voice, audio, video, and sensing data online
Log Aggregator	Collect application data, logs, messages, events, etc.
RDB Aggregator	Collecting structured data from relational database

데이터 수집 과정은 아주 중요한 부분으로 본 논문에서는 파이썬(Python)을 이용하여 수집하였다.

3.2 Tokenization

토큰화(tokenization) 과정은 데이터를 임의의 문자열로 변환하는 과정을 의미한다.

본 논문에서는 원시 데이터를 주제와 부주제별로 크롤링을 통해 수집한 후, 전체 데이터를 하나의 파일로 통합하였다. 이후 사용자 단어 사전과 금지어, 불용어 및 변환 사전을 구성 하는데 있어 언어적 특성과 상관없이 적용될 수 있는 엔그램 언어 모델(n-gram language model) 방식을 적용하였다. 여기서 엔그램(n-gram)의 n에 해당하는 단어의 수에 따라 하나일 때는 유니그램(unigram)과 두 개일 때는 바이그램(bigram), 세 개일 때는 트리그램(trigram) 등을 적용하여 토큰화를 수행하였다. 수행 결과를 가지고 형태소 분석을 하여 분석 데이터 셋을 생성하였다.

3.3 Characteristic analysis

형태소 분석을 한 데이터 셋을 이용해 특성 분석을 진행하였다. 특성 분석에서는 단어의 빈도수와 역문서 빈도수를 곱한 TF-IDF(Term Frequency - Inverse Document Frequency)[08] 데이터 값을 생성하고 이 값을 룰 기반의 데이터 셋으로 구성하였으며 대표 단어를 추출하는 방법을 이용하여 머신러닝을 위한 문서 단어 행렬(DTM:Document Term Matrix) 데이터를 구성하였다[18].

TF-IDF 알고리즘은 토큰화된 단어의 빈도수를 나타내는 단어의 빈도수(tf : Term Frequency) 값이며 문서의 빈도를 나타내는 문서 빈도(df : Document Frequency) 값에 역수를 취한 역문서 빈도(idf : Inverse Document Frequency)를 나타낸 값이다. TF-IDF 값은 단어의 빈도수와 역문서의 빈도 값을 곱한 값이며 수식은 아래의 1)과 같다.

- 단어빈도수 : tf(토큰화된 단어가 나타난수(t), 전체단어(d))
- 역문서빈도수 : idf(토큰화된 단어가 포함된 문서수(df), 특정 단어(t))

$$= \log \frac{\text{전체문서수}(N)}{1+\text{토큰화된단어가포함된문서수}(df)}$$

1) $tf-idf = tf_{t,d} * \log \frac{N}{1+df_t}$

2) $tf-idf = tf(t,d) * idf(d,t)$

문서 단어 행렬은 특정 단어의 빈도수를 행렬로 표현하는 방식으로 시행하였다.

3.4 Analytical Model

분석 모델 단계에서는 분류 모델(Classification Model)에 정확도 검증을 위하여 소프트맥스 회귀(Softmax Regression)와 서포트 벡터 머신(SVM : Support Vector Machine), 그리고 랜덤 포레스트의 분석 모델을 적용하였다.

소프트맥스 회귀는 각 클래스의 확률을 추정하여 이중 가장 높은 확률을 가진 클래스를 선택하는 방법으로 알고리즘은 아래와 같다.

$$S(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{k} e^{x_j}}$$

서포트 벡터 머신은 인공지능이나 데이터 마이닝 기법 등에 쓰이는 대표적인 알고리즘 중 하나이며 분류율이 좋은 분류 알고리즘 중 하나이다.

랜덤 포레스트[15]는 다수의 의사결정트리를 학습하는 수학적 방법론으로 높은 정확성과 간편하고 빠른 학습 효과를 얻을 수 있다. R에서 적용한 랜덤 포레스트 알고리즘은 아래와 같다.

$$\tilde{f}(x) = \frac{1}{n}\sum_{i=1}^{n} f(x, x_{iC})$$

위에 서술한 분석 모델들을 적용하여 분석을 진행하였으며 그중 가장 정확도가 높은 알고리즘을 선택하여 비정형 데이터 분석의 최종 결론을 도출하였다.

IV. Implementation and Testing

4.1 Collecting thesis data and creating a database

디비피아 홈페이지에서 6,002건의 논문 데이터를 Fig. 4.와 같이 파이썬을 이용하여 크롤링하고 Fig. 5.와 같이 수집된 데이터를 논문의 주제 및 부 주제별 중복 데이터를 배제하고 내용이 없는 데이터에 대한 제거 작업을 수행한 후 3,628건의 데이터로 분석 작업을 수행하였다.

```
for search_keyword in keywords:
    print('search_keywords::', search_keyword)
    while True:
        try:
            search_urls = f'http://www.dbpia.co.kr/search/topSearch?startCount=0&collection=ALL&range=A&searchField' \
                          f'=ALL&sort=RANK&query={search_keyword}&srchOption=*&includeAr=false'
            paper_urls = get_paper_urls(search_urls)  # search_keyword를 검색한 결과로 나타난 paper url 30개의 list
            result_dict = {'topic': [], 'topic_keyword': [], 'paper_title': [], 'paper_source': [],
                           'paper_keywords': []}
            break
        except:
            continue

    for paper_url in paper_urls:
        print('paper_url::', paper_url)
        while True:  # 오류가 나면 재시도
            try:
                result = crawl_paper_page(paper_url)  # crauling 한다.
                break
            except:
                continue

        result_dict['topic'].append(topic)
        result_dict['topic_keyword'].append(search_keyword)
        result_dict['paper_title'].append(result[0])
        result_dict['paper_source'].append(result[1])
        result_dict['paper_keywords'].append(result[2])
```

Fig. 4. Crawled Python Source Code

Fig. 5. Collection Dataset

4.2 Word tokenization and data Preconfiguration

4.2.1 Modeling and Data dictionary Building

알고리즘 구현을 위하여 Fig. 6.과 같이 데이터 저장을 위한 모델링을 수행하였다.

코엔엘피(KoNLP)의 데이터 사전을 이용하여 데이터베이스를 생성하는 작업을 수행하였다. 형태소 분석을 위하여 엔그램 언어 모델 중 유니그램과 바이그램 그리고 트리그램을 활용해 단어의 토큰화 과정을 진행하였으며 카이스트의 9품사 분류 체계를 적용하여 명사만 추출하였다[02].

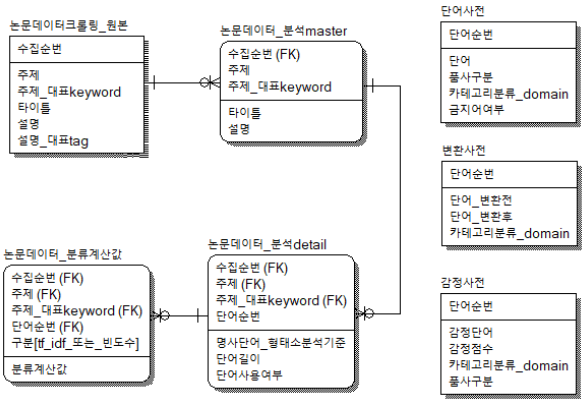

Fig. 6. Modeling Processing

4.2.2 Word dictionary Calibration Method

단어 교정 변환 기준으로는 Fig. 7.과 같이 불용어와 금지어를 기준으로 제거하고 변환어 작업을 진행하였다. 사용 기준으로는 실체가 명확한 것(물리적으로 존재), 업무에 명확한 단어로 진행하였다. 변환 기준으로는 형용사 + 명사 형태의 어절, 명사 + 어미 형태의 어절을 진행하였고, 금지어 기준으로는 불특정 다수가 공용으로 사용 가능한 단어 제거, 추상적인 단어 제거, 형용사/ 부사 형태의 단어를 제거하는 방법으로 진행하였다.

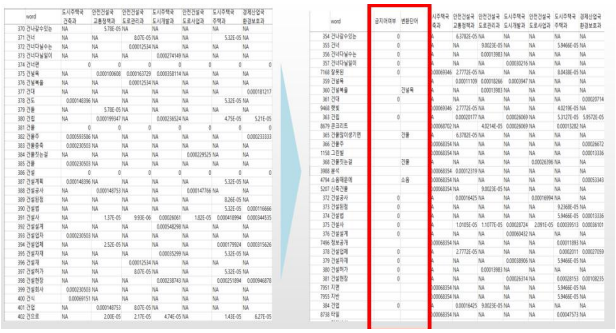

Fig. 7. The way words are pre-calibrated.

4.3 Analyze attributes by classification and organize analysis data sets

TF-IDF 데이터를 생성하고 학습 데이터와 와이 값(y value)에 해당하는 데이터를 결합하여 최종 분석 데이터 셋을 생성하였으며, TF-IDF 값을 이용하여 문서 단어 행렬(DTM) 데이터를 구성하였다.

4.4 Classification model validation

4.4.1 Perform analysis model

멀티 카테고리(Multi Category) 선택을 통한 주요 단어 빈도 기준으로 활용하였고, 데이터에 대한 특성을 이해하고 탐색하기 위해 Fig. 8.과 같이 R을 통해 시각화를 진행하였다.

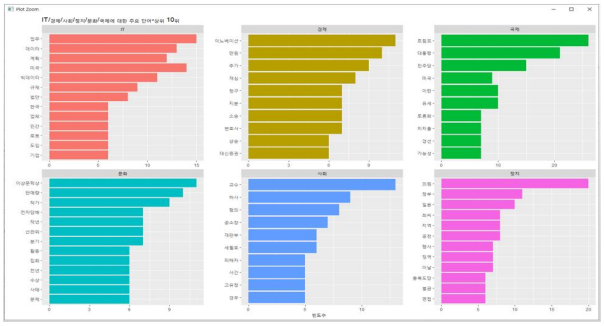

Fig. 8. Basic Navigation – Visualization

4.4.2 Verification of accuracy

모델링의 정확도 검증을 위하여 데이터에 대한 시각화 작업을 진행하였고, 그 중 룰 기반의 TF-IDF 값을 활용한 분류가 가장 정확도가 높게 나왔으며, 이를 Fig. 9.에 시각화를 활용하여 표현하였다.

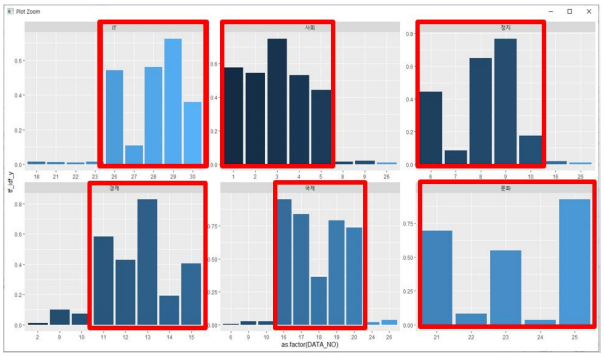

Fig. 9. Classification result of tf-idf and rule-based utilization classification.

4.5 Analysis Result

```
>
> class_fore_relt <- bind_rows(class_fore_relt, data.frame('예측모델' = '의사결정나무',
+                                                          'train 정확도' = rpart_train_relt,
+                                                          'test 정확도' = rpart_test_relt))
>
> class_fore_relt
                예측모델 train.정확도 test.정확도
1           RandomForest    1.0000000   1.0000000
2 Support Vector Machine    0.8619403   0.8846154
3           의사결정나무    0.8395522   0.8846154
>
```

Fig. 10. Analysis Result

Table 7. Measurement Value

Analysis Model	Analysis Result	Test Analysis Result
RandomForest	1.00	1.00
Support Vector Machine	0.86	0.88
Decision Tree	0.83	0.88

분석 모델에서 정확한 분류 기법을 찾기 위하여 랜덤 포레스트, 서포트 벡터 머신. 의사결정트리 분석의 알고리즘을 적용하였고, 이 중 랜덤 포레스트는 Fig. 10. 및 Table. 7. 과 같이 분석 결과와 테스트 분석 결과가 모두 1.00 이나와 정확하게 분류되었다는 것을 알 수 있었다.

이는 애초에 사용된 데이터가 디비피아의 요약과 주제어를 사용하여 분류한 결과로 이미 잘 분류된 데이터에 대하여 모델을 적용하였기에 가능한 결과로 보여진다.

V. Conclusions

빅데이터 시대에는 데이터에 대한 가치 분석을 위하여 비정형 데이터에 대한 분석 및 처리가 더욱 많이 요구되고 있다.

본 논문에서는 빅데이터 시대에서 많이 요구되는 비정형 데이터에 대한 분류를 위하여 분석 모델을 설계하고 이를 검증하였다.

이를 위하여 데이터는 디비피아의 논문 요약과 주제어, 그리고 부주제 어를 크롤링하여 전처리 작업 등을 통해 중복 및 배제 데이터들에 대한 정제 작업을 진행하였다. 또한, 코엔엘피의 데이터 사전을 이용해 데이터베이스를 생성하였으며, 형태소 분석을 통하여 단어의 토큰화 과정을 수행하였고, 카이스트의 9 품사 분류 체계를 이용해 명사를 추출하였다.

추출된 명사를 이용하여 TF-IDF 값을 생성하였으며, 학습 데이터와 와이 값을 결합하여 분석 데이터 셋을 생성하였다. 생성된 분석 데이터 셋에 세 가지 분석 알고리즘을 적용하여 분류의 적정성을 측정하였다.

비록 본 논문에서는 사용된 데이터가 이미 잘 분류된 논문 관련 데이터이기에 100% 정확도를 보였지만 다른 데이터 셋에서는 이렇게 정확하게 나오지는 않을 것으로 예측된다.

향후 본 연구의 내용은 민원 분석 및 텍스트 관련 분석 등으로 활용 범위를 넓혀 나갈 수 있을 것으로 기대한다.

ACKNOWLEDGEMENT

This study was supported by Gangneung-Wonju National University.

REFERENCES

[1] Barnett, T. P., and R. Preisendorfer. (1987). “Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis.”.

《Monthly Weather Review 115》
[2] Key-Sun Choi, Young S. Han, Young G. Han, Oh W. Kwon, KAIST tree bank project for Korean: Present and future development, In Proceedings of the International Workshop on Sharable Natural Language Resources, pp. 7-14, 1994
[3] Cho Taeho "Concepts and Applications of Text Mining“, Journal of scientific & technological knowledge infrastructure no.5, 2001, pp.76 - 85
[4] Leo Breiman (2001). “Random Forests”. 《Machine Learning》 45 (1): 5−32. doi:10.1023/A:1010933404324
[5] “Regression analysis”《Encyclopedia of Mathematics》. Springer-Verlag. 2001. ISBN 978-1-55608-010-4.
[6] Choi YunJeong, Park SeungSoo "Interplay of Text Mining and Data Mining for Classifying Web Contents" The Korea Society for Cognitive Science 13(3), 33-46, 2002
[7] Hsu, Daniel, Sham M. Kakade, and Tong Zhang (2008). “A spectral algorithm for learning hidden markov models.”. 《arXiv preprint arXiv:0811.4413》
[8] Manning, C. D.; Raghavan, P.; Schutze, H. 《Introduction to Information Retrieval》. Cambridge University Press. 100-123. ISBN 9780521865715. 2008 Scoring, term weighting, and the vector space model
[9] Douglas, Laney. " 3D Data Management: Controlling Data Volume, Velocity and Variety ." Gartner. Retrieved February 6, 2001
[10] Beom Jiin, Choi Sungjong, "Bigdata use cases and implications", CEO Focus Vol. 312, 2013
[11] EunSoon You, GunHee, Choi, SeungHoon Kim "Study on Extraction of Keywords Using TF-IDF and Text Structure of Novels", Journal of The Korea Society of Computer and Information Vol. 20, No. 2, February 2015
[12] Mary Meeker's 2016 internet trends report
[13] Kamiński, B.; Jakubczyk, M.; Szufel, P. (2017). “A framework for sensitivity analysis of decision trees”. 《Central European Journal of Operations Research》. doi:10.1007/s10100-017-0479-6
[14] Park Jooseok "A Comparative Study of Big Data, Open Data, and My Data", Korea Bigdata Society, 41-46, No 3, Vol. 23, 2018
[15] Liaw, Andy March 25, 2018. “Documentation for R package randomForest”
[16] Kim HyunJong, Lee TaiHun, Ryu SeungEui, Kim NaRang "A Study on Text Mining Methods to Analyze Civil Complaints: Structured Association Analysis", Journal of the Korea Industrial Information Systems Research Vol. 23 No. 3, 2018.6
[17] Cho ByungSun "A Comparative Study on Requirements Analysis Techniques using Natural Language Processing and Machine Learning", Ajou Univ. 2020.
[18] Bryan Bischof. Higher order co-occurrence tensors for hypergraphs via face-splitting. Published 15 February, 2020, Mathematics, Computer Science, ArXiv
[19] HyunJin Yeo “Mobile Commerce Brand Identity Strategy by SNS Text mining”, Journal of The Korea Society of Computer and Information, Vol. 25 No. 10, October 2020
[20] Hello data science - www.hellodatascience.com Jinyoung Kim
[21] Data collection - www.dbguide.net KOREA Data Agency

Authors

SungJin Kim is a visiting professor with the Department of Computer Science and Engineering at Myongji College. He received as B.S., M.S. degree in Computer Science from Halla University Wonju, Korea in 2013, 2015 and currently receiving Ph.D. in data science part in multimedia engineering at GangNeung-Wonju National University. He is currently a specialist in the Bigdata Consulting and worked for 27 years in the IT Business. His research interests include AI, IoT, BigData, advanced data analyst, business process architecture.

NakJin Choi is team leader at Korea Expressway Corporation. He received as M.S. degree in Computer Science from Korea University, Korea in 2006 and currently receiving Ph.D. in data science part in multimedia engineering at GangNeung-Wonju National University. He has been in is information system engineer and data analysis for 33 years with the IT department at Korea Expressway Corporation. He is currently an team leader and data analysis expert at Korea Expressway Corporation. His research interests include AI, IoT, advanced data analyst, and platform.

JunDong Lee is a professor with the Department of Multimedia Engineering at GangNeung-Wonju National University. He received as B.S., M.S., and Ph.D. degree in Computer Science from HongIk Unversity, Seoul, Korea in 1990, 1993, and 2001, respectively. His research interests include programming language, IoT, and platform.