빅데이터 처리 기술을 활용한 비정형데이터 분석 모델링 구축

김정훈O, 김성진*, 권기열*, 주다혜*, 오재용*, 이준동*
*강릉원주대학교 멀티미디어공학과,
O강릉원주대학교 멀티미디어공학과
e-mail: wjdgns778@naver.comO, tonyksj@naver.com*, qqqq4632@naver.com*, jdh12188@naver.com*, ojy5251@naver.com*, jlee@gwnu.ac.kr*

Building Modeling for Unstructured Data Analysis Using Big Data Processing Technology

Jung-Hoon KimO, Sung-Jin-Kim*, Gi-Yeol-Kwon*,
Da-Hye-Ju*, Jae-Yong-Oh*, Jun-Dong Lee*
*Dept. of Multimedia Engineering, GangNeungWonju University,
ODept. of Multimedia Engineering, GangNeungWonju University

● 요 약 ●

기업 및 기관 데이터는 워드프로세서, 프레젠테이션, 이메일, open api, 엑셀, XML, JSON 등과 같은 텍스트 기반의 비정형 데이터로 구성되어 있습니다. 텍스트 마이닝(Textmining)을 통해서 자연어 처리 및 기계학습 등의 기술을 이용하여 정보의 추출부터 요약 · 분류 · 군집 · 연관도 분석 등의 과정을 수행울 진행한다. 다양한 시각화 데이터를 보여줄 수 있는 다양한 모델 구축을 진행한 후 민원 신청 내용을 분석 및 변환 작업을 진행한다. 본 논문은 AI 기술과 빅데이터를 활용하여 민원을 분석을 하여 알맞은 부서에 민원을 자동으로 할당해 주는 기술을 다룬다.

키워드: 빅데이터(Big Data), 데이터(Data), 분석(Analysis), 처리(Processing), 시각화(Visualization), 텍스트 마이닝(Textmining), 모델링(Modeling)

I. Introduction

매년 민원신청 데이터는 개별적 이유 등으로 무분별하게 늘어나는 추세이다. 하지만 실제로 민원신청 데이터를 확인해 보면 민원 내용이 없고, 주제만 있는 경우가 있고, 민원 내용에 맞춤법, 오타, 띄어쓰기, 다양한 은어 사용 등 여러 가지 잘못된 단어를 사용하고 있었고 민원 내용을 일일이 확인하여 분류하는 형태로 진행이 되고 있었다.

본 논문은 총 52개의 부서 민원 내용을 수집을 하여, 내용 분석 및 분류 처리 기술을 활용해 민원 내용에 해당되는 부서에 맞게 분류를 진행해 주는 내용을 다룬다.

II. Preliminaries

본 논문에서 처리하는 빅데이터는 기존의 데이터와 속성이 달라 데이터 수집·저장·처리·분석·표현하는 데 새로운 방법들이 필요하다.

Fig. 1. 은 빅데이터를 처리하는 과정을 크게 데이터의 생성·수집·저장·처리·분석·표현의 과정으로 분류한 것이다[1].

Fig. 1. Big Data Processing Process

III. The Proposed Scheme

1. 민원 데이터 수집 및 분석 활용 Dataset

광주 시청 홈페이지에서 1,676건의 VOC 데이터를 crawling으로 작업한 후 Fig. 2. 와 같이 수집하여 분석에 착수(총 52개의 부서에 대한 데이터 수집) 했다.

Fig. 2. Utilization Dataset

2. 분석 방법론

본 논문의 데이터를 분석하는 방법으로는 TEXT MINING 기법을 활용하였고, 과제 수행 및 알고리즘 개발을 위하여 데이터 분석 방법에 특성 이해를 위해 Fig.3와 같은 그림으로 표현해 보았다.

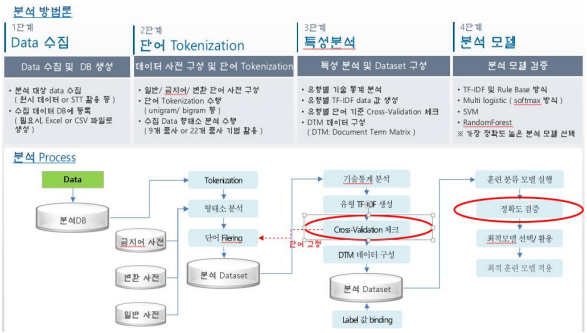

Fig. 3. Textmining Processing

3. Data 사전 구성 및 교정 방식

- Modeling 및 Data 사전 구축

과제 및 알고리즘 구현을 위하여 Fig. 4. 와 같이 데이터 저장을 위한 Moedling을 수행하였고, KoNLP에서 사전을 받아서 DB에 생성시키는 작업을 수행하였고, R을 활용해 Data 분할 작업 수행한 뒤 단어가 1자 이상인 것만 발취하였다.

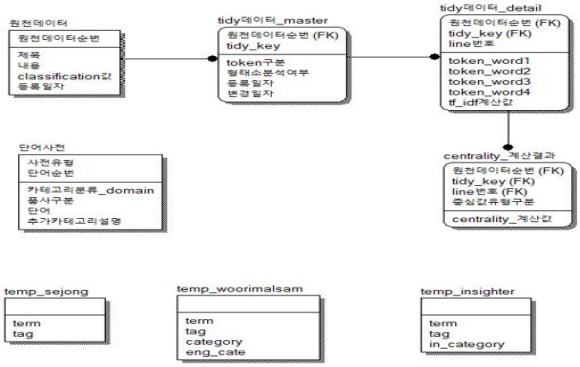

Fig. 4. Modeling Processing

- 단어 사전 교정 방식

단어 교정 변환 기준으로는 Fig. 5.와 같이 사용, 변환, 금지어 기준으로 진행하였고, 사용 기준으로는 실체가 명확한 것(물리적으로 존재), 업무에 명확한 단어로 진행하였다. 변환 기준으로는 형용사 + 명사 형태의 어절, 명사 + 어미 형태의 어절을 진행하였고, 금지어 기준으로는 불특정한 부서에서도 사용 가능한 단어 제거, 추상적인 단어 제거, 형용사/ 부사 형태의 단어를 제거하는 방법으로 진행하였다.

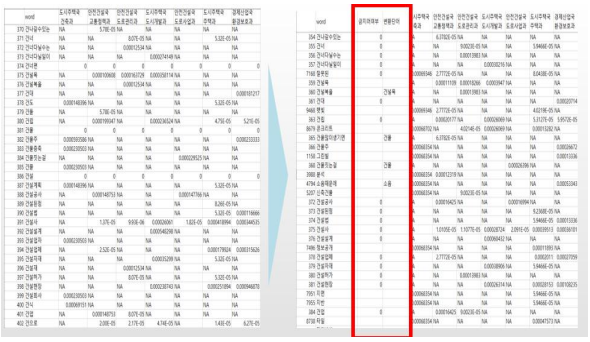

Fig. 5. The way words are pre-calibrated.

4. Data 탐색 및 검증 방법

- Data 탐색

Multi Category 선택을 통한 주요 단어 빈도 기준으로 활용히였고, 데이터에 대한 특성을 이해 및 탐색하기 위해 Fig. 6. 과같이 R을 통해 시각화를 진행하였다.

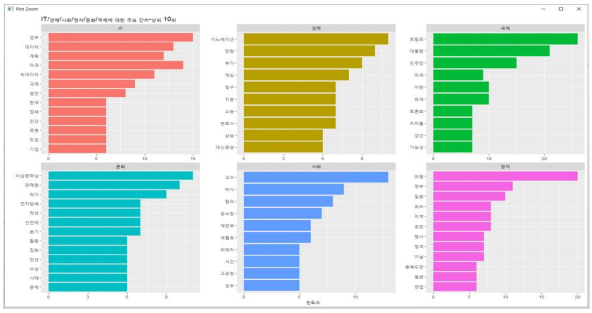

Fig. 6. Basic Navigation - Visualization

- 검증 방법

모델링에 데이터를 활용하여 정확도 검증을 위해 R을 활용하여 시각화 작업을 진행하였고, 다양한 시각화 분석 방법 중에 가장 정확도가 높게 나왔던 Fig.7 와 같이 tf_idf, Rule 기반으로 활용해서 진행하였다.

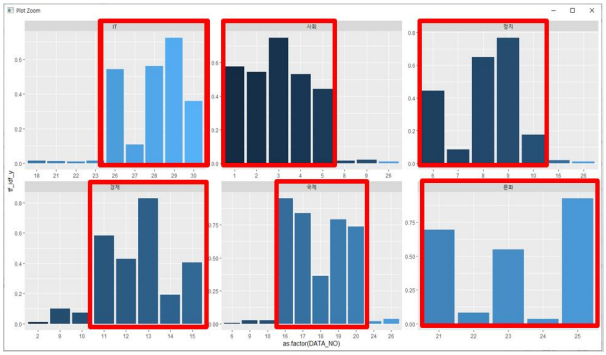

Fig. 7. Classification result of tf_idf and rule-based utilization classification.

IV. Conclusions

분석 모델의 특징으로 다양한 분석 모델 적용해본 결과 단어 사전, 시각화 작업으로 진행하였는데 단어 사전으로는 분류 모델에 대한 분석 방법은 일반적 사전 + 업무 사전을 기준으로 형태소 분석 수행이 가장 적절했고, 형태소 분석에 대한 문제 해결은 단어 변환으로 해결하였다. 시각화 작업으로는 여러 가지의 분류 모델 중에 가장 정확도가 높은 것을 선택적으로 활용하였다(tf_dif, Rule 방식 활용하였고 Rule 방식에서는 확률 값을 기준으로 한 오 분류에 내용 파악이 가능했다.

ACKNOWLEDGMENT

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 한국연구재단 – 현장맞춤형 이공계 인재양성 지원사업의 지원을 받아 수행된 연구임 (NRF-2017H1D8A1029391).

REFERENCES

[1] Hanbit Academy, "IT CookBook, Big Data Computing Technology," pp. 17, June 2014.
[2] Multicampus, "Big Data Distributed Parallel Processing System and Introduction to Analysis" pp. 35 ~ 51, February 2020.
[3] 잔재미코딩, https://www.fun-coding.org/crawl_basic2.html
[4] Multicampus, "Big Data Distributed Parallel Processing System and Introduction to Analysis" pp. 319 ~ 413, February 2020.
[5] 인투더데이터, https://intothedata.com/02.scholar_*category/text_mining/*
[6] 인투더데이터, https://intothedata.com/40.r_language/
[7] 인투더데이터, https://intothedata.com/02.scholar_category/text_mining/tfidf/
[8] DEEPPLAY, https://3months.tistory.com/343