Journal of the Korea Industrial Information Systems Research Vol. 23 No. 3, Jun. 2018 :13-24
http://dx.doi.org./10.9723/jksiis.2018.23.3.013 ISSN:1229-3741

민원 분석을 위한 텍스트 마이닝 기법 연구: 계층적 연관성 분석

(A Study on Text Mining Methods to Analyze Civil Complaints: Structured Association Analysis)

김 현 종[1], 이 태 헌[2], 유 승 의[3], 김 나 랑[4]*
(Kim HyunJong, Lee TaiHun, Ryu SeungEui, and Kim NaRang)

요 약 정부 및 공공기관에 있어 시민의 직접적인 요구사항이 담겨 있는 민원은 정책 개발을 위한 중요한 데이터로 활용이 가능하다. 그러나 민원 데이터는 비정형 텍스트로 작성되어 있는 특성으로 인해 일반적인 텍스트 마이닝 기법으로는 시민의 요구사항을 정확히 도출하기 어려웠다. 이에 본 연구에서는 민원 데이터 분석을 위한 텍스트 마이닝 기법을 개선하여, 시민의 요구사항을 도출할 수 있는 방법을 제시하고자 하였다. 새로운 텍스트 마이닝 기법은 공기어구조맵의 원리에 착안하여 연관성 분석을 2단계로 실시하여 핵심주제어를 기반으로 1차 연관 단어 와 2차 연관 단어로 구조화 하였다. 분석을 위해 2016년 1년간 부산시 민원게시판에 올라온 3004건을 활용하였다. 분석 결과는 빈도수와 핵심주제어를 가지고 연관성 분석만으로는 찾을 수 없었던 민원 상의 문제를 본연구에서 제시한 계층적 연관성 분석을 이용하여 시민의 요구사항을 더욱 정확하게 파악할 수 있었다. 본 연구는 민원 데이터에서 시민의 요구사항을 도출하기 용이한 방법을 제안하였다는 학문적 기여점이 있으며, 행정기관에서 민원 데이터를 통해 정책 개발에 활용할 수 있다는 실무적 기여점이 있다.

핵심주제어 : 텍스트 마이닝, 계층적 연관성 분석, 전자민원, 민원 분석

Abstract For government and public institutions, civil complaints containing direct requirements of citizens can be utilized as important data in developing policies. However, it is difficult to draw accurate requirements using text mining methods since the nature of the complaint text is unstructured. In this study, a new method is proposed that draws the exact requirements of citizens, improving the previous text mining in analyzing the data of civil complaints. The new text-mining method is based on the principle of Co-Occurrences Structure

* Corresponding Author : whitecoral@hanmail.net
+ 이 논문 또는 저서는 2015년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2015S1A3A2046781).
Manuscript received May 31, 2018 / revised June 18 ,2018 / accepted June 27, 2018
1) 동아대학교 경영정보학과, 제1저자
2) 동아대학교 산학협력단, 제2저자
3) 동아대학교 경영정보학과, 제3저자
4) 동아대학교 경영정보학과, 교신저자

Map, and it is structured by two-step association analysis, so that it consists of the first-order related word and a second-order related word based on the core subject word. For the analysis, 3,004 cases posted on the electronic bulletin board of Busan City for the year 2016 are used. This study's academic contribution suggests a method deriving the requirements of citizens from the civil affairs data. As a practical contribution, it also enables policy development using civil service data.

Key Words : Text Mining, Structured Association Analysis, Civil Complaints, Complaint Analysis

1. 서 론

정부 및 공공기관에서는 행정서비스의 향상을 위해 많은 노력을 기울이고 있다. 이 중에서도 민원업무는 행정기관과 국민과의 직접적인 소통 역할을 담당하고 있다. 과거 정부기관에서는 개인 및 단체의 개별적인 행정서비스를 제공하기 위한 수단으로 민원처리의 효율 제고에 중점을 두었다. 이후 인터넷 등의 정보통신기술의 발달로 정부 2.0이 등장하면서 양방향 소통과 맞춤형 행정서비스의 제공을 중요시하기 시작하면서 개인의 행정업무뿐만 아니라 정책과 행정서비스의 불만사항 및 개선사항의 제안이 증가하고 있다. 더 나아가 전자정부와 정부 3.0 시대에서는 빅데이터가 사회적 관심사로 부각되면서 민원 데이터의 분석에 대한 관심이 높아지고 있으며 민원처리의 관리체계 및 시스템에 대한 변화가 시작되었다. 이에 빅데이터 기반의 중앙 및 지방 정부의 민원을 분석하는 연구도 증가하고 있다.

민원 분석은 수요자 중심의 정책 개발과 정책 품질의 향상을 위해서 중요하다. 그러나 민원과 같은 비정형 텍스트 분석의 어려움으로 행정기관에서 활용이 부족한 편이다. 특히 민원의 경우 맞춤법, 표준어의 사용에 있어서 언론 기사나 논문 등에 비해 부정확한 경우가 많고, 제기되는 주제와 내용이 너무 다양하여 데이터 분석만으로 민원 데이터에서 의미 있는 결과를 찾아내는 것이 어렵다. 그리고 일반적으로 많이 사용하는 키워드 분석, 소셜 네트워크 분석만으로는 정확한 민원의 요구사항을 파악하는 것에 한계가 있다.

이에 본 연구는 텍스트 마이닝 기법을 개선하여 전자민원 데이터의 새로운 분석 방법을 제안하고자 한다. 이 방법을 사용하여 부산시 전자민원 데이터에 대한 텍스트 마이닝의 연관성 분석을 2단계에 걸쳐 계층적으로 분석하였으며, 분석 결과를 통해 새로운 방법론에 대한 검증과 시민의 요구사항을 도출하고자 한다.

2. 이론적 배경

2.1 전자민원

민원은 '국민이 행정기관에 대하여 처분 등 특정한 행위를 요구하는 것'으로 정의된다[1]. 즉, 민원은 국민이 행정기관에 답변 및 행위를 요청하는 의사표현을 통칭하는 개념으로 국민이 간편하게 이용할 수 있는 행정구제 수단이다. 또한 민원은 개인 및 단체의 특정 행정 업무를 처리하는 것이지만 주민들의 요구사항과 불만 및 개선사항 등 주민의 의견을 제안할 수 있는 창구 역할도 한다.

전자민원은 전자정부를 기반으로 민원행정서비스를 기관에 직접 방문하지 않고도 업무를 온라인 방식으로 처리하는 방식이다[2]. 정부 2.0의 등장과 전자정부법 등을 통해 온라인 방식의 전자민원시스템이 마련되어 민원 처리의 효율성 향상과 양방향 의사소통체계를 확보하고자 하였다. 민원처리의 기본방침은 행정차치부 등의 법령으로 정해져 있으며 국민신문고와 같은 시스템의 운영 및 관리는 국민권위위원회에서 관장하고 있다. 현재까지의 민원처리는 민원 업무의 효율성과 공정성 등에 중점을 두고 있어 민원 정보의 분석에 대한 개념과 기능이 부족하다[3].

민원을 통해 주민들은 개인 생활 불편, 행정서비스의 개선, 정책 제안 등 다양한 요구사항을 제시하고 있다. 이러한 민원 정보는 공공서비스 및 정책 개발의 계획단계에서 활용될 수 있으며 시민의 요구사항에 대한 행정 대응성을 향상시킬 수 있다[4]. 전자민원은 효율적인 민원처리 외에 시민의 다양한 의견을 수집, 분석할 수 있어 정책결정과정에 중요한 자료로 활용될 수 있다. 이는 정책 개발에 시민이 참여하는 하나의 방법이라고 할 수 있다.

2.2 텍스트 마이닝 분석 방법

텍스트 마이닝은 문자(text) 기반의 데이터에서 새로운 정보를 찾기 위한 방법이다. 빅데이터 분석 기법 중 인터넷 및 소셜 미디어 등에서 발생하는 비정형 텍스트 데이터를 자연어 처리 및 문서 처리 기술을 활용하여 가치 있는 정보를 추출하는 것을 목적으로 한다. 텍스트 마이닝은 아래 Table 1과 같이 다양한 방법들이 있다.

핵심 주제어 분석은 문서 전체에서 자주 등장하는 단어가 중요한 핵심이 될 수 있기 때문에 가장 우선적으로 파악하는 분석 방법이고 단순하게 출현 빈도에 따라 순위를 부여하는 방법과 TF-IDF 등의 방법을 사용하여 전체 문서가 아니라 특정 문서 내에서 중요도를 수치화 하여 핵심어를 추출할 수 있다[5]. 단어 연관성 분석과 동시 출현 단어 분석은 같은 문서에 동시에 등장하는 단어들이 가지는 연관성과 패턴을 수치화 하거나 시각화하는 방법이다. 연관성 분석은 장바구니 분석 등 마케팅을 비롯한 다양한 분야에서 활용되고 있다[6]. 이슈 토픽 분석은 문서 집합에서 자주 등장하는 유사한 단어들을 통해 잠재된 주제를 발견하는 방법이다. 소셜 네트워크 분석은 전일욱 외[7]의 연구와 같이 텍스트 마이닝 기법은 아니지만 텍스트 마이닝 기법으로 추출된 단어들을 네트워크로 구성하여 각 단어들의 연결과 연관성을 수치화하고 시각화해 주는 방법이다[8].

3. 단어 연관성의 계층적 분석 방법론

일반적인 텍스트 마이닝의 분석 과정은 Fig. 1과 같이 수집된 텍스트 데이터를 분석 목적에 맞게 전처리를 실시한다[9]. 전처리 시에는 문서 및 문장을 품사 단위로 정리하는 형태소 분석이 이루어진다. 이후 핵심 주제어를 분석하는 경우가 많고 이때 주로 등장하는 단어의 빈도를 기준으로 주요 키워드를 추출하고 키워드의 특성을 정의하거나 분석한다. 이어서 연관성 분석, 이슈 토픽 분석, 감성 분석 텍스트 마이닝의 여러 방법 등을 연구 목적에 맞게 선택하여 실시하게 된다.

민원 분석에서도 텍스트 마이닝의 다양한 방법을 활용할 수 있다. 단어의 빈도는 시민이 많이 언급된 단어를 통해 주요 관심사를 파악할 수 있는 핵심 주제어 분석이 있다. 단순히 많이 등장

Table 1 Text-Mining Analysis Method

Method	Description
Core Subject Analysis	A method to extract frequently mentioned keywords from a specific group of documents and sort them according to the frequency
Association analysis	A method to computes the correlation between frequently appearing words in a document and quantify the level of association
Co-appearance words analysis	Analysis to know the pattern of co-appearance in a document among keywords
Emotion analysis	Estimation method to understand the emotional status of writers from documents
Issue/Topic Analysis	A method to grasp automatically and extract specific topics or issues from a document
Social Network Analysis	A method to understand the connection and association among words in the composed word network of a document

하는 것만으로는 문제를 정확하게 파악할 수 없으며, 일상적으로 많이 사용되는 단어가 높은 빈도를 차지할 수 있다.

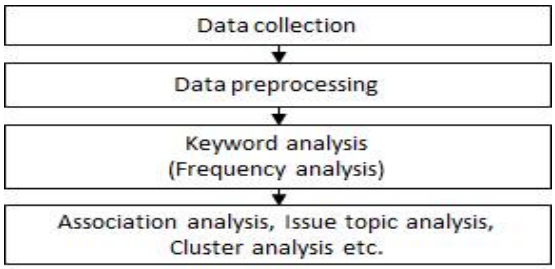

Fig. 1 Process of General Text Mining Analysis

그리고 토픽 분석은 문서의 주제를 찾아내는 방법으로 문서 내의 단어를 통해 문서 집단에 숨겨진 주제와 잠재적 의미 구조를 분석할 수 있다. 토픽 분석을 통해 민원에서 주요 이슈를 구성하는 단어들과 이를 대표하는 주제를 추출 할 수 있다[10]. 이 분석은 빈도 분석보다 정확하게 민원에서 이슈가 되는 주제를 정확하게 추출할 수 있지만, 핵심 주제어와 마찬가지로 토픽 분석만으로는 시민의 요구사항을 구체적으로 파악하기에는 부족하다. 그리고 동시 출현 단어 분석 및 단어 연관성 분석은 문서 내에 동시에 출현하는 단어들을 분석하는 방법으로 단어들의 구조적, 상관관계를 파악할 수 있다[11]. 연관성 분석은 현윤진 외[6]의 연구와 같이 신문기사에서 관심 단어들의 연관성 규칙을 밝히고 연관성 높은 단어를 도출할 수 있다. 민원 분석에서도 연관성 분석 통해 민원의 내용에서 연관성 높은 단어를 추출하여 민원 내용을 파악할 수 있다. 그러나 관심도가 높은 주제어 와 연관 단어만을 추출하게 되면 주요 민원에 대한 일반적이고 개략적인 내용만을 파악할 수밖에 없다. 그 외 감성분석, 군집분석, 소셜 네트워크 분석 등을 통해 민원을 분석하고 있으나 대부분 앞서 설명한 방법들의 수준에서 이루어진다.

이와 같이 일반적인 텍스트 마이닝의 분석 방법만으로는 전체 민원 데이터에서 시민의 요구사항을 파악하기가 어렵다. 이를 개선하기 위한 방법으로 연관성 분석을 여러 단계로 실시하여 구조화 하는 방법을 구상하였다. 관련 연구로 정하영외[4]는 공기어구조맵(Co-Occurrences Structure Map)을 작성하는 방법으로 주요 키워드와 동시에 출현하는 제1층 공기어를 추출하고, 이 공기어의 제2층 공기어를 다시 추출하여 주요 키워드에 대한 공기어구조맵을 작성함으로써 핵심 주제어에 대한 연관 단어를 폭넓게 구조화 할 수 있다고 주장하였다.

이에 민원 데이터를 통해 시민의 요구사항을 더욱 정확하고 용이하게 분석하기 위한 방법으로 핵심 주제어의 연관성 분석을 계층적으로 구조화하는 Fig. 2와 같은 과정의 방법을 고안하였다.

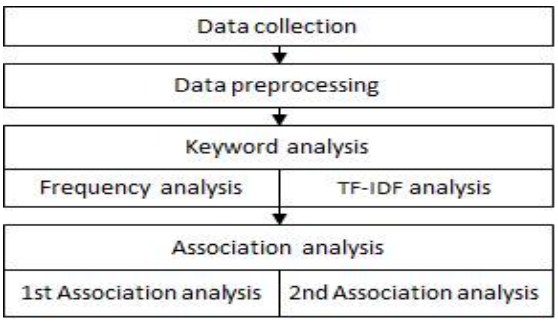

Fig. 2 Process of Hierarchical Association Analysis

계층적 연관성 분석 과정은 데이터 수집 및 전처리와 핵심주제어 분석은 일반적인 텍스트 마이닝 기법과 동일하다. 그러나 핵심 주제어 분석에서 빈도 분석과 함께 TF-IDF 분석을 추가하여 문서 내에서 단어 중요도가 높은 단어를 핵심 주제어 선정에 활용하도록 하였다. 연관 분석은 핵심 주제어를 1차 연관성 분석하여 추출한 연관 단어에 대해 2차 연관성 분석을 실시하여 2단계에 걸쳐 연관 단어를 추출한다. 이를 통해 추출된 1, 2단계의 연관 단어를 핵심 주제어 - 1단계 연관 단어 - 2단계 연관 단어로 계층화 하는 것이다. 이렇게 계층적 구조의 연관 단어들을 연결하여 민원의 내용을 좀 더 정확히 파악할 수 있다.

4. 전자 민원 분석 사례

전자민원 분석의 과정과 사용된 R 프로그램의 패키지는 Fig. 3과 같다.

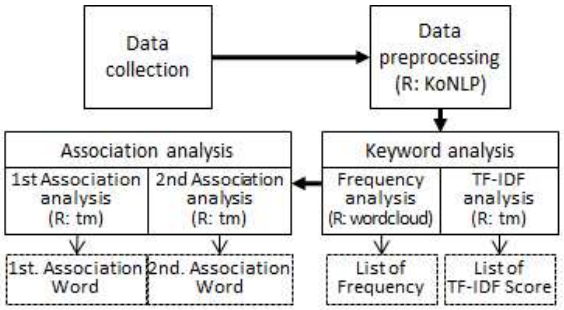

Fig. 3 Process of Electronic Civil Service Analysis

4.1 데이터 수집 및 전처리

본 연구에서 부산시 홈페이지 민원 게시판의 데이터를 이용하였다. 이 민원 게시판에는 연 1만 건 이상의 민원이 접수되고 있으며, 민원 신청자에 선택에 따라 공개 또는 비공개로 등록이 되고 있다. 비공개 게시물의 경우 시 담당자 외에 열람이 되지 않기 때문 수집할 수 없었으며 개인적인 민원이 주로 이루어지므로 수집 대상에서 제외하였다. 이에 본 연구에서는 민원 게시물 중에서 수집 가능하며, 공공적인 부분에 관한 민원이 많은 공개 게시물을 수집하였다. 수집된 민원은 2016년 1년간 등록된 3004건의 데이터이다.

수집된 민원 데이터는 자연어로 작성된 비정형 텍스트이기 때문에 분석을 위해 자연어처리가 필요하다. 먼저 문서형태의 자료를 품사 단위로 구분하여 필요한 데이터를 추출하는 전처리를 위해 통계 프로그램 R의 KoNLP 패키지를 이용하여 형태소 분석을 실시하여 민원 내용에서 명사 단위로 키워드를 추출하였다.

4.2 핵심 주제어 분석

본 연구에서는 민원에서 다루어지는 주요 키워드를 분석하기 위해 데이터 전처리 과정에서 추출한 명사 단어를 이용하여 핵심 주제 분석을 실시하였다. 핵심 주제 분석은 일반적인 빈도 분석과 TF-IDF((Term Frequency Inverse Document Frequency) 분석을 활용하였다. 빈도 분석은 전체 문서에서 등장하는 빈도로 단어의 중요성을 파악할 수 있다. 그러나 전체 문서에서 특정 단어가 많이 등장하는 것만으로 중요도가 높다고 할 수는 없다. 이를 보완하기 위해 TF-IDF 분석을 실시하였다. TF-IDF 분석은 문서 중에 단어의 중요성을 판정하는 방법이다[12]. 문서에 등장하는 단어의 빈도인 TF와 단어가 등장하는 문서의 빈도의 역수인 IDF를 곱한 값으로 단어의 중요도를 판정한다. TF-IDF 값이 높다는 것은 특정한 단어가 문서에서 중요하다는 것을 의미한다. 문서 전체에서 흔하게 사용되는 단어가 아니라 특정 문서에 중요도가 높은 단어이다. 이를 통해 문서 내에서 중요한 단어인 토픽을 찾을 수 있다[4].

핵심 주제어 분석을 위해 전처리한 민원 데이터를 통계 프로그램 R의 워드클라우드 패키지와 tm 패키지를 사용하였다. 추출된 명사의 빈도수는 아래 Table 2와 같다.

전체 민원에서 빈도가 높은 단어들 중에는 부산, 답변, 시민 등과 같이 민원 내용에 관용적으로 사용되는 단어를 비롯하여 저와, 누구, 그것 등과 같이 의미가 불분명하거나 무의미한 단어를 제외하여야 한다.

빈도 순으로 나타난 핵심 주제어로는 '버스'가 가장 높은 빈도를 보이고 있으며 관련 단어로 보이는 '노선', '기사', '정류장' 등도 높게 나타났다. 그리고 교통과 관련된 '차량', '신호', '정차' 등도 높게 나타났다. 그외 '아파트', '공사', '설치', '안전' 등의 단어가 높게 나타났다.

다음으로 TF-IDF 분석을 통해 개별 문서에서 중요도가 높은 단어를 분석하였다. TF-IDF 분석 결과는 Table 3과 같다.

TF-IDF 분석 결과 '버스', '정류장', '노선', '택시', '차량' 등과 같이 교통과 관련된 단어들이 높은 값을 보였다. 그 외 '단속', '아파트', '공사', '설치' 등의 단어가 높게 나타났다.

빈도 분석과 TF-IDF 분석을 비교한 결과는 Table 4와 같으며 두 결과의 상위 100위 이내의 단어 중 무의미한 단어를 제외한 단어들의 순위를 비교한 것이다. '버스', '정류장', '노선', '이용', '불편', ' 차량', '단속', '기사', '아파트', '공사', '승객', '관리', '운행', '설치', 등의 단어가 공통적으로 높은 값을 나타내고 있다. 이들 단어는 전체 문서에서도 높은 빈도를 가지며 문서별로 중요도가 높은 단어로 핵심 주제어로 분류할 수 있다. 그리고 '택시',

‘도로’, ‘간격’, ‘교통’, ‘주차’, ‘출퇴근’, ‘서류’, ‘차선’ 등의 단어는 TF-IDF 순위가 빈도에 비해 높은 단어로 빈도수는 적은 편이나 문서 내에서 중요도가 높은 단어인 것으로 나타났다. 그리고 ‘정차’, ‘불법’, ‘안전’, ‘조치’, ‘해운대’, ‘문제’, ‘신호’ 등의 단어는 빈도수가 TF-IDF에 비해 높은 순위로 나타나 여러 문서에 등장하며 문서 내에서는 중요도가 적지만 전체적으로 많이 언급된 단어들이다.

Table 4에서 나타난 단어 중 상당수는 교통과 관련되어 있으며 그중에서도 버스와 관련된 것으로 추정되는 ‘정류장’, ‘노선’ 등의 단어는 제외하고 핵심 주제어를 선정하였다. 핵심 주제어는 버스, 불편, 택시, 차량, 단속, 도로, 아파트, 공사, 주차, 설치 등 10가지를 선정하였다.

Table 2 Frequency Analysis Result

No.	Word	Freq.	No.	Word	Freq.	No.	Word	Freq.	No.	Word	Freq.
1	Bus	3459	21	To	669	41	Haeundae	414	61	Reason	348
2	Busan	2364	22	Race	596	42	Facility	411	62	Degree	348
3	Think	1331	23	We	586	43	Telephone	407	63	Official	346
4	They	1246	24	Inconvenience	573	44	Passenger	406	64	Us	342
5	Citizen	1215	25	Install	571	45	Doing	405	65	If	336
6	Use	1172	26	Station	557	46	Use	403	66	Area	336
7	Time	1159	27	Management	528	47	Request	403	67	Solution	333
8	Person	1106	28	Illegal	517	48	Myungji	399	68	Guidance	332
9	Route	1033	29	Not	514	49	Because	390	69	Manager	331
10	Busan	1029	30	Measure	506	50	Taxi	388	70	Matters	327
11	Complaints	948	31	Signal	504	51	Female	383	71	Subway	324
12	Article	933	32	Safety	491	52	Need	380	72	Parking	323
13	Citizen	864	33	Stop	483	53	Operation	378	73	Town	315
14	Apartment	817	34	Thank	473	54	You	367	74	Road	313
15	Construction	809	35	Situation	471	55	Contents	364	75	More than	312
16	Ward Office	783	36	Occation	468	56	Progress	361	76	Education	311
17	Answer	774	37	Business	466	57	Process	360	77	Danger	304
18	Intermittent	758	38	Market	463	58	Change	359	78	Dispatch	303
19	Problem	737	39	City hall	450	59	Children	355	79	Relation	299
20	Vehicle	720	40	Confirm	424	60	Occur	350	80	Street	298

Table 3 TF-IDF Analysis Result

No.	Word	TF-IDF score	No.	Word	TF-IDF score	No.	Word	TF-IDF score	No.	Word	TF-IDF score
1	Bus	1381.59	21	Article	206.59	41	Race	144.46	61	Effort	114.52
2	Time	839.85	22	Apartment	197.19	42	City	143.92	62	Center	112.98
3	Station	423.49	23	Interval	195.36	43	Install	143.35	63	Curious	112.76
4	Route	418.60	24	Citizen	192.73	44	I	139.11	64	Yacht	111.45
5	Use	376.46	25	Think	187.94	45	Enrollment	137.45	65	Driving	110.59
6	Inconvenience	339.97	26	Construction	187.80	46	Answer	135.06	66	Go to work	109.68
7	Taxi	315.73	27	Place	183.28	47	Notification	134.73	67	Ward Office	109.68
8	Complaints	298.81	28	Passenger	174.78	48	Office	134.73	68	Illegal	109.37
9	Citizen	287.87	29	Change	174.70	49	Lane	133.73	69	Safety	109.37
10	That	269.91	30	Manager	174.61	50	Stop	133.73	70	Diesel	107.79
11	Time	262.80	31	Attachment	170.88	51	Bridge	132.96	71	Intersection	107.79
12	Vehicle	261.61	32	Thank	165.47	52	Telephone	130.07	72	Persons	105.64
13	Minute	249.93	33	Use	165.42	53	Time	129.72	73	Direction	105.45
14	Intermittent	246.56	34	Official	160.54	54	Way	125.96	74	Measure	105.39
15	Moorage	234.25	35	Traffic	158.56	55	City Hall	125.80	75	Call	103.88
16	Person	233.72	36	And	152.22	56	Place	124.49	76	Side	102.78
17	Car	230.80	37	Parking	151.33	57	Front	124.06	77	City policy	102.74
18	Number	226.95	38	Commute	148.20	58	Guidance	123.06	78	Money	102.67
19	Roughness	223.08	39	Document	145.92	59	Request	121.06	79	Permission	100.64
20	Road	216.83	40	Management	145.63	60	Issued	117.71	80	Degree	99.02

Table 4 Comparison of TF-IDF and Frequency Results of Key Words

Word	TF-IDF ranking	Frequency ranking	Word	TF-IDF ranking	Frequency ranking
Bus	1	1	Install	44	25
Station	3	26	Lane	50	107
Route	4	9	Stop	51	33
Use	5	6	Guidance	59	68
Inconvenience	6	24	Driving	66	93
Taxi	7	50	Illegal	69	28
Vehicle	13	20	Safety	70	32
Intermittent	15	18	Diesel	71	-
Road	21	74	Measure	75	30
Article	22	12	Need	83	52
Apartment	23	14	Haeundae	84	41
Interval	24	132	Card	88	131
Corporation	27	15	Problem	92	19
Passenger	29	44	Subway	98	71
Traffic	36	110	Signal	-	31
Parking	38	72	Children	-	59
Commute	39	163	Town	-	73
Document	40	-	Education	-	76
Management	41	27	Danger	-	77
Race	42	22	Dispatch	-	78

4.3 연관성 분석

핵심 주제어 분석을 통해 추출한 단어의 출현 빈도와 중요도만으로는 의미를 찾기가 힘들다. 문서의 맥락적 의미를 찾기 위해서는 주요 단어들의 구조적 형태와 연결 관계를 파악하여야 한다. 이를 위해 단어 간의 상호 관계를 분석하는 연관성 분석을 실시한다. 연관성 분석은 동시에 출현하는 단어의 특정 순서로 발생하는 확률로 나타내며 이를 통해 단어의 의미적 접근성 및 상호의존성을 발견할 수 있다. 사용되는 분석 방법으로는 동시출현 단어 분석과 단어 연관성 분석이 있다.

연관성 분석을 위해 통계 프로그램 R의 tm 패키지 패키지를 사용하였다. 핵심 주제어로 선정된 단어로 연관성 분석을 실시한 결과는 아래 Table 5와 같다.

핵심 주제어의 연관성이 높은 단어만으로는 민원에서 제기되는 의미를 파악하기 힘들다. 분석을 통해 나타난 연관 단어는 상식적인 수준에서 유추 가능하며, 빈도 조사에서도 어느 정도 나타난다. 그러나 시민들이 민원을 통해 제기한 내용을 더 자세히 파악하기에는 한계가 있다.

핵심 주제어의 1차 연관 단어를 2단계 연관성 분석을 실시하여 2차 연관 단어를 추출하였다. 이를 통해 핵심 주제어에 연관되는 단어들의 연결을 폭 넓게 조사할 수 있으며 민원에서 제기하는 내용에 가까운 단어의 배열을 찾을 수 있다. Table 6은 핵심 주제어 중 가장 비중이 높은 '버스'의 연관 단어에 대한 연관성 분석을 실시한 결과이다. 2차 연관 단어의 결과도 Table 5와 같이 연관계수가 있으나 삭제하고 높은 순으로 표기하였다.

2단계에 걸친 연관어 분석을 통해 핵심주제어-1차 연관어-2차 연관어 순으로 나열할 경우 Fig. 4와 같이 민원에 대한 내용을 파악할 수 있었다. 연결된 단어의 조합은 버스-정류장-정차, 버스-정류장-버스알림기계, 버스-기사-친절(불쾌), 버스-기사-급출발, 버스-배차-간격, 버스-배차-증차, 버스-정차-정위치, 버스-운행-최적, 버스-운전-급출발, 버스-출발-앞차, 버스-간격-배차, 버스-노선-연장 등과 같이 나타난다. 결과에서 보여지듯이 1차 연관 단어들과 2차 연관 단어들 사이에 또 다른 연관 관계가 나타나기 때문에 3단계 이상의 연관 관계도 유추할 수 있다.

Table 5 Association Analysis Result of Keyword

Keyword	Association analysis result									
Bus	Station (0.44)	Driver (0.41)	Downtown (0.37)	Public transport (0.34)	Time (0.33)	Allocation (0.32)	Stop (0.3)	Race (0.28)	Passenger (0.28)	Driving (0.26)
	Town (0.26)	Start (0.25)	Interval (0.23)	Stop (0.23)	Use (0.22)	Penalty (0.22)	Route (0.21)	Go to Work (0.21)	Arrive (0.2)	Myungji (0.2)
Taxi	Indirect tax (0.42)	Earned Income Tax (0.42)	Social order (0.42)	Reorganization (0.42)	Government subsidy (0.42)	Skyrocketing (0.42)	Hantang (0.42)	Taxi Fee (0.41)	individual (0.36)	Reduce car (0.36)
	Time (0.34)	Corporation (0.33)	VAT (0.33)	Infant (0.33)	Boarding area (0.33)	Qualifications System (0.33)	Throw up (0.32)	Driver (0.31)	Cacao (0.31)	National tax (0.3)
Vehicle	Parking (0.28)	Traction (0.23)	Equivalent (0.22)	Financial pressure (0.22)	Financial (0.22)	Final goal (0.22)	Illegal parking (0.21)	Hybrid (0.21)	Notice (0.2)	Eco (0.2)
	Order (0.2)	Instruction sheet (0.2)	Bookstore (0.2)	Signal (0.19)	Illegal parking (0.19)	Passage (0.19)	Truck (0.19)	Lane (0.18)	Penalty (0.18)	Exemption (0.18)
Road	Deduction (0.46)	Myung rwun dong (0.45)	Open (0.45)	Building law (0.45)	Boundary survey (0.45)	Public sewer (0.45)	Stop construction (0.45)	Reconstruction (0.45)	Directly (0.45)	Reply (0.45)
	Auditor chief (0.45)	Results notification (0.45)	Resonance pole (0.45)	National ombudsman (0.45)	Nationalization (0.45)	auditing office (0.45)	Road drawing (0.45)	Next door (0.45)	Hazardous trees (0.45)	Road deduction (0.45)
Parking	Parking lot (0.45)	Line (0.42)	Black (0.42)	Yunsan market (0.42)	Parking problem (0.35)	Cavitation (0.35)	Balanced development (0.35)	Underdeveloped area (0.35)	Partnership (0.35)	Model (0.35)
	Attentive (0.35)	Rough (0.35)	Incentive policy (0.35)	Electric pole (0.31)	Van (0.31)	Parking fee (0.3)	Humiliation (0.3)	Light car (0.28)	General vehicle (0.27)	Residence (0.26)
Apartment	Other revenue (0.36)	Council (0.33)	Unprotected (0.32)	Acceleration (0.32)	Difficult (0.32)	City (0.32)	Block (0.32)	School (0.32)	Ramp (0.32)	Elasticity (0.32)
	Security service (0.31)	Announcement (0.31)	Construction work (0.31)	Advertising Agencies (0.31)	Installer (0.31)	Sticker (0.31)	Elevator (0.31)	Re-bid (0.31)	Move in (0.3)	Construction company (0.29)
Construction	Green core (0.4)	Stop (0.39)	Criminal disposition (0.39)	Concrete (0.38)	Original (0.38)	Resumption (0.38)	Expansion construction (0.38)	Excitement (0.38)	Audit name (0.38)	Construction industry (0.38)
	Machinery (0.38)	Unlicensed (0.38)	Specification (0.38)	Elected office (0.38)	Defeat (0.38)	Voluntary resignation (0.38)	Vote (0.38)	Construction site (0.37)	License (0.37)	Boiler (0.37)
Install	Material (0.28)	Bus information (0.27)	Financial circumstances (0.27)	Gradual (0.27)	Convenience (0.27)	Corrugated board (0.27)	Vandalism (0.27)	Protruding type (0.27)	By color (0.27)	Forced processing (0.22)
	Lightweight (0.22)	Functional (0.22)	Unification (0.22)	Logo (0.22)	Finishing material (0.22)	Customized (0.22)	Name plate (0.22)	Humidity (0.22)	Lighting facility (0.22)	Solar power (0.22)
Inconvenience	Use (0.28)	Route (0.21)	Bus (0.2)	Convenient (0.19)	Register system (0.19)	Insult (0.19)	Front door (0.19)	Card (0.19)	Student (0.19)	Presbyopia (0.19)
	Registration system (0.19)	Disproof (0.19)	Fraudulent use (0.19)	Switch (0.19)	Get off (0.19)	Operation (0.18)	Traffic congestion (0.18)	Usage pattern (0.18)	Race (0.17)	Direct (0.17)
crack down	Police (0.4)	Overload vehicle (0.4)	Worker's name (0.4)	Chief officer (0.4)	Interdiction service (0.4)	Leader (0.4)	Staying up all night (0.4)	Cotton bat (0.4)	Driver (0.4)	Bank interest (0.4)
	Punisher (0.4)	Prize (0.4)	Omission (0.37)	Parking violation (0.32)	Animal (0.31)	Eradication (0.28)	Sale (0.27)	Slaughter (0.26)	Dog market (0.25)	Official (0.25)

Table 6 Association Analysis Result of 'Bus'

Keyword	1st. Association Word	2nd. Association Word
Bus	Station	Blue, Stop, Rear car, Window, Outside, Grumbling, Empty, First time, Bus search, Bus notification machine, Criminal activity, Thunderbolt, Bamboo, Naver, Flexibility, Forbidden city, Distance, Driver, Collar
	Driver	Bus, Driving, Passenger, Taxi, Boarding, Kindness, Destination, Mood, Company, Quick start, Riding, Back door, Get off, Petulance, Personality education, Traffic charges, Discomfort, Stop, Discipline, Charge
	Time	Fines, Family business, Physiological phenomena, Joint, Bus, Inefficiency, Fines, Union, Meal, Quasi-public system, Traffic Violations, Downtown, Allocation, Go to work, Go Home, Unreasonable, Lunch, Commute, Starting, Interval
	Allocation	Interval, Route, Bus, Race, Increase, Shorten, Transfer, Operating Distance, Time, Bus reduction, Time allowed, Algebraic Effect, Reduction effect, Visitor, Traffic condition, Commute, Extension, Net increase, Total amount, Reduction
	Stop	Regular Position, conductor, Bus, pocket type, Station, Driving Lane, passenger, Advancement, Stop, driver, blue, Get off the bus, Boarding, Guide signs, customs, accusations, anger, Rear magnet, employer, blog
	Race	Route, Direct, Bus route, Traffic congestion, Ban-Yu dong, Requires time, Usage pattern, Jea-Song dong, Lane diagram, Regular, Plan, Optimal, Centum city, Normally, User, Integrated, Tolerance, Allocation, Shorten
	Passenger	Officers, Drivers, Substitutes, Back Door, Front door, Turn, Inconvenience, Bus, Riding, Document, Commonality, Route type, Get off, Low, Simplicity, Line, Stop, Section, Use, Card
	Driving	Driver, Downtown, Example, Pride, Fines, Family Business, CEOs, Physiological phenomena, Joint, Overseas study, Operator, Stewardess, Union, Marginalized, Traffic violations, Quick start, Transportation business, Compliance, Quasi-public system
	Start	Front car, Bus, Terminal point, Instability, Procedures, Speedboat, Step, Driver, Terminal, Tightness, Time, Service spirit, Recovery, Business, Arrival, Back door, Exact time, Getting off, Passengers, Waiting
	Interval	Allocation, Vehicle reduction, Route, Allowable time, Shortening, Operating distance, Transfer, Increase, Race, Reduction effect, Rear car, Wage workers, Traffic conditions, Money, Bus, Total amount system, Designation, Turn, Commute, Demand
	Route	Race, Direct, Ban-Yu dong, Usage patterns, Requires time, Bus routes, Integrated, Regular, Demand, Centum city, Utilization realities, Extended, Optimal, Section, Am, Plan, Diesel, Jeugn-Gwan, Visitors, Abolition

Table 7 Analysis Result and Civil Requirement

Analysis Result	Civil Requirement
Bus - Station - Stop - Regular Position	Inconvenience of entry/exit due to inaccurate stop of the bus at Stations
Bus(Taxi) - Driver - Kindness - Unpleasant Attitude	Unpleasant experience caused by unkind bus or taxi drivers training is required
Bus - Allocation - Interval- Bus Increase	Interval of bus allocation is too sparse. More buses are required.
Bus - Drive - Route - Optimization (Extension)	The bus route should be optimized or extended (Banyeo, Jeonggwan)
Bus - Drive - Driver -Quick start	The quick start of bus drivers has risks of accidents.

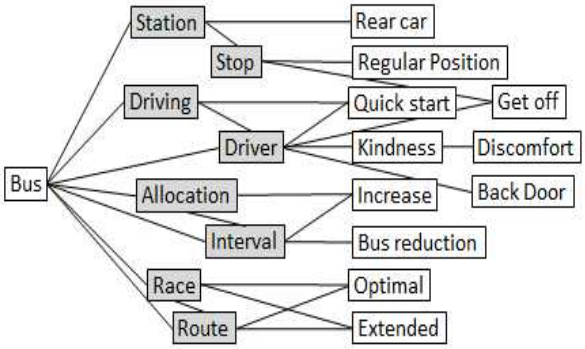

Fig. 4 Connect the Associated Word

4.4 시민 요구사항 도출 및 시사점

계층적 연관성 분석 결과에서 나타난 연관 단어들의 연결을 통해 버스 관련 시민의 대표적인 요구사항을 아래 Table 7과 같이 정리할 수 있다.

버스와 연관성이 높은 10여개의 단어들은 서로 어느 정도 연관성을 가지고 있어 쉽게 문제를 파악되어 지는 것 같으나 이 단어들만으로는 해석하기 힘든 부분이 존재한다, 예를 들어 '정류장과 '정차'는 연관성이 높으며 의미가 상통하기 때문에 "정류장의 정차에 문제가 있다"는 것을 찾을 수 있지만 구체적인 어떤 문제나 원인을 제시해주지는 못한다. 그러나 '정차'의 연관 단어인 '정위치'가 발견됨으로써 정위치에 정차하지 않는 문제임을 파악할 수 있다. 또한 '배차'와 '간격'도 "배차 간격에 문제가 있다"고 하는 단편적인 문제 외에 '증차'에 대한 요구사항과 '감차'라는 원인을 비롯하여 '출퇴근', '교통여건'에 맞는 배차 간격 조정을 요구한다는 것을 파악할 수 있었다.

요약하면, 기존의 민원 데이터를 비롯한 비정형 텍스트 분석에서 주로 빈도를 이용하여 핵심 주제어를 선정하기 때문에 빈도가 낮은 단어나 핵심 주제어와 연관성이 낮은 단어는 발견되지 않는 문제가 있었다. 이러한 문제는 본 연구에서 실시한 분석 결과와 같이 핵심 주제어에서부터 1, 2차 연관어의 각 단어를 이어 주는 것만으로도 시민의 요구사항이 어렵지 않게 파악되어 질 수 있다. Table 7에서는 결과의 예를 들기 위한 5가지만을 제시하였지만 관련 문제에 대한 다양한 조합을 통해 폭 넓은 문제 파악과 분석이 가능할 것이다.

5. 결 론

본 연구는 텍스트 마이닝을 활용하여 전자민원 데이터를 분석하고 정책을 도출하기 위한 텍스트 마이닝 의 새로운 방법론을 제안하고 있다. 기존 연구에서 주로 사용되던 빈도 분석과 연관성 분석 방법을 개선하여 민원에서 시민의 요구사항을 파악하고 정책 개발에 필요한 의제를 도출하였다. 특히 텍스트 마이닝의 연관성 분석을 2 단계로 실시하여 계층적인 구조의 연관 단어를 추출하고 이를 통해 민원의 요구사항을 파악하였다. 이를 위해 선행연구를 통해 기존 텍스트 마이닝을 개선하기 위한 방안을 모색하였으며, 이를 검증하기 위해 부산시의 2016년 전자민원 3004건에 대하여 핵심주제어를 분석하고 이중 가장 높은 빈도를 나타내는 '버스'의 연관 단어 10개에 대해 2차 연관성 분석을 실시하였다. 그 결과 핵심주제어-1차 연관 단어-2차 연관 단어의 연결을 통해 시민들의 버스와 관련한 요구사항을 파악할 수 있었다.

본 연구는 비정형 텍스트 분석에서 주로 사용되는 텍스트 마이닝 방법인 빈도 분석과 연관성 분석만으로는 분석 결과에서 내용을 파악하기 힘든 점을 개선하기 위해 연관성 분석을 여러 단계를 걸쳐 계층적인 구조로 연관 단어 추출하는 방법을 제시하고 있다, 이 방법을 통해 일반적인 텍스트 마이닝 방법에서 빈도수가 적거나 핵심주제어와의 연관성이 낮아 발견하기 어려웠던 민원의 원인과 시민의 요구사항을 추출할 수 있다. 이와 같이 추출된 원인과 시민의 요구사항 통해 사회의 문제점과 정책 과제를 구체화시킬 수 있다.

학문적 기여점으로 본 연구에서 제시된 텍스트 마이닝의 개선된 방법과 절차를 통해 비정형 텍스트의 분석 결과를 향상시킬 수 있으며, 텍스트 마이닝을 비롯한 민원 분석 및 행정분야의 후속연구를 위한 자료로 활용할 수 있다는 점을 들 수 있다. 또한 행정기관에서 민원 데이터의

분석을 위한 분석 기법으로 활용하여 보다 쉽게 민원 내용을 파악하고 이를 통해 정책과제의 도출과 방안을 제안할 수 있다는 실무적인 기여점이 있다.

본 연구에서 제시한 계층적 연관성 분석 방법은 과거 연구된 적이 없는 방법으로 정하영 외 [4]의 공기어구조맵에 착안하여 분석함으로써 의미있는 확실한 이론적 근거가 부족하다는 한계점이 있다. 향후 연구에서는 다년간의 자료와 전체 민원 데이터를 대상으로 여러 분야에 대한 요구사항을 분석하여 민원의 시계열적 특성 및 구조적 특성을 파악할 수 있는 연구와 함께 2단계 계층구종의 타당성을 뒷받침할 수 있는 이론적 연구가 필요하다.

References

[1] e-Peple, https://www.epeople.go.kr/jsp/user/pc/ cvreq/UPcCvreqInfo.paid, Accessed 23 March 2018.

[2] Park, G. G. and Jung J. H., "A Study on the Determining Factors of the On-Line Civil Administrative Service into Civil Satisfaction," The Korean Journal of Local Government Studies, Vol. 16 No. 4, pp. 359-380, 2012.

[3] Chi, W. J., Sim, J. S., Nam, S. W. and Her, J. S., "A Study on Development Method of Civil Data Analysis based on Big Data," POP Consulting, 2015.

[4] Jeong, H. Y., Lee, T. H., Hong, S. G., "A Copus Analysis of Electronic Petitions For Improving the Responsiveness of Public Services: Forcusing on Busan Petiton," The Korean Journal of Local Government Studies Vol. 21, No. 1, pp. 423-436, 2017.

[5] Cho, T. I., "Spatiotemporal Characteristics Analysis of Complaints on Officially Assessed Land Price by Big Data Mining," Graduate School of Incheon National University, 2015.

[6] Huyn, Y. J., Kim, J. S., Jeong, J. W., Yun, S. M. and Lee, M. S., "Text Mining on Internet-news Regarding Climate Change and Food," Journal of the Korean Data And Information Science Sociaty, Vol. 26, No. 2, pp. 419-427, 2015.

[7] Jeon, I. W., Jun, O. J., Choi, M. Y., Kim, H. S. and Chung, J. H., "Characteristics of Civil Complaints to a Local Government based on Social Network Analysis: Focused on Cheonan City E-Bulletin Board (Allso 365)," Journal of Regional Studies, Vol. 25, No. 2, pp. 117-141, 2017.

[8] Park, J. S., Hong, S. G. and Kim, N. R., "A Development Plan for Co-creation-based Smart City through the Trend Analysis of Internet of Things," Journal of the Korea Society Industrial Information System, Vol. 21, No. 4, pp. 67-78, 2016.

[9] Lee, J. H. and Lee, H. G., "A Study on Unstructured Text Mining Algorithm through R Programming Based on Data Dictionary," Journal of the Korea Society Industrial Information System, Vol. 20, No. 2, pp. 113-124, 2015.

[10] Park, J. S., Hong, S. G. and Kim J. W., "A Study on Science Technology Trend and Prediction Using Topic Modeling," Journal of the Korea Society Industrial Information System, Vol. 22, No. 4, pp. 19-28, 2017.

[11] Park, J. H. and, Pi, S. Y., "A Study on WT-Algorithm for Effective Reduction of Association Rules," Journal of the Korea Society Industrial Information System, Vol. 20, No. 5, pp. 61-69, 2015.

[12] Suchman, M. C., "Managing Legitimacy: Strategic and Institutional Approaches," Academy of Management Review, Vol. 20, No. 3, pp. 571-610, 1995.

김 현 종 (Kim HyunJong)

- 정회원
- 동아대학교화학과 이학사
- 동아대학교경영정보학과 경영학석사
- 동아대학교경영정보학과 경영학 박사 수료

- 관심분야 : 정보시스템, 빅데이터, 텍스트 마이닝, Co-creation

이 태 헌 (Lee TaiHun)

- 정회원
- 동아대학교도시계획학과 공학사
- 리츠매이칸대학교정책과학연구학 이학석사
- 리츠매이칸대학교정책과학연구학 이학박사

- 관심분야 : 지역혁신, 마을만들기, 지역정책, 지리정보시스템

유 승 의 (Ryu SeungEui)

- 정회원
- 동명정보대학교경영정보학과 경영학사
- Texas A&M International University 경영학석사

- University of Texas at El Paso 경영학박사
- 관심분야 : 경영정보시스템, 빅데이터, 오피니언 마이닝

김 나 랑 (Kim NaRang)

- 정회원
- 부산대학교문헌정보학과 공학사
- 동아대학교경영정보학과 경영학석사
- 동아대학교경영정보학과 경영학 박사

- 관심분야 : 지역혁신, Co-creation