

공공서비스의 수요대응성 향상을 위한 전자민원 코퍼스분석: 부산시 사례를 중심으로*

정 하 영**
이 태 현***
홍 순 구****

국문요약

시민들의 공적관심사에 대한 다양한 내용분석이 시도되어 왔으나 대부분이 조사자가 설계한 설문조사를 통해서 공적관심사를 집약하는 방법이었다. 이러한 방법은 조사자의 주관에 의존하기 때문에 조사자가 달라질 경우 분석결과가 판이하게 달라질 수 있으며 데이터를 집약하기 까지 많은 시간과 노력이 요구된다. 이에 본 연구에서는 문제탐색과 의사결정을 위한 빅데이터 과학의 일환인 코퍼스 언어학 기반의 내용분석수법 (Corpus Linguistic Based Content Analysis) 을 응용하여, 민원데이터 속에 내포되어 있는 시민들의 실질적인 공적관심사(Public Concern)를 객관적으로 기술하고 자동적으로 집약하는 공적관심사 평가수법(Public Concern Assessment Methodology)을 개발하는 것을 목표로 한다. 본 연구의 목적 달성을 위해 1) TFIDF(Term Frequency Inverse Document Frequency) 수법을 이용한 주요민원 분석방법론과 2) 공기어(Co-Occurrences) 추출과 공기어구조맵을 시각화하여 주요민원에 대한 공통인식 및 의견대립 분석방법론을 제안하고, 문제점 및 정책과제를 검토하였다. 이를 위해 전자민원 창구 “부산시에 바란다”에 2015년 1월 1일 - 2015년 6월 31일까지 공개 접수된 민원데이터에 대한 코퍼스분석을 실시하였다. 본 연구는 다양한 민원인의 개별적 문맥을 고려한 키워드 추출, 공통민원 탐색, 상충된 민원 탐색, 애로민원 탐색 등 정책입안자에게 신속하고 정확하게 민원내용을 파악할 수 있는 방법론을 제안하는데 공헌하고 있다.

주제어: 지역혁신, 부산시 전자민원, 코퍼스 분석, 텍스트 마이닝, 공적관심사(Public Concerns)

* 이 논문은 2015년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2015S1A3A2046781).

** 제1저자

*** 교신저자

**** 공동저자

I. 서론

정부는 최근 전자민원시스템을 도입하여 공공서비스에 관한 민원을 신속하게 집약할 수 있는 기틀을 마련하였다. 그러나 방대한 민원내용을 읽고, 시민들의 실질적인 공적관심사(public concern)를 집약하는 데에는 천문학적인 시간과 비용이 소요되기 때문에 사람이 처리하기에는 사실상 불가능하다. 이러한 문제를 해결하기 위한 방법으로 코퍼스 언어학 기반의 내용분석수법(Corpus Linguistic Based Content Analysis)이 있다. 이를 응용하여 “부산 해수담수화플랜트사업”을 대상으로 미디어간에 존재하는 인식차이를 객관화하기 위한 선행연구가 진행된바 있다(이태현 외, 2017). 이 연구에서는 해수담수화 플랜트 가동이라는 사안에 대한 공적관심사를 집약하기 위하여 시계열로 핵심어를 추출하는 자동화된 오피니언 마이닝 기법을 구축하여 대립이나 갈등을 해소할 수 있는 단서를 제공하는 것을 목적으로 하였다. 하지만 미디어 텍스트는 시민들의 여론을 형성하는 것을 목적으로 한다는 점에서 실질적인 시민들의 공적관심사를 반영하기에는 부족함이 존재한다.

또한, Jeong et al.(2008)은 TFIDF(Term Frequency-Inverse Document Frequency), T-VSM(Term-based Vector Space Machine) 및 MDS(Multi-Dimensional Scaling)를 주요 방법으로 하는 포함한 코퍼스 담화분석방법을 토대로 공공사업의 사회적 적절성을 판단하기 위한 방법으로 주로 사업주체간의 회의, 주민협의회 회의, NGO단체와 지역행정과의 회의 등 작은 규모의 토론내용을 계량화하는 연구를 진행하였다. 본 연구는 보다 광범위한 내용을 포함하는 광역시의 민원을 분석하기 위한 연구로 확장하였다. 이와 같이 맥락의 파악이 불가능한 대량의 텍스트로부터 누가 시도하더라도 동일한 결과물을 낼 수 있는 계량적이고 파악가능한 정보를 추출하고자 하는 연구는 빅데이터 분석, 오피니언 마이닝 등의 분야에서 주로 다루어지고 있다. 불특정 다수의 분산된 의견으로 이루어진 비정형 데이터로부터 평판분석을 통한 마케팅 분야의 활용(서지훈, 2014), 뉴스 콘텐츠를 분석함으로써 주가예측정보로 활용하는 지능형투자의사결정 모형개발(김동영, 2014), 쇼핑물 상품평의 검색효율을 높이기 위한 정보검색방법론에 관한 연구(Yang, 2009; Yune, 2010)등 다양한 분야에서 비정형텍스트데이터로부터 의미 있는 정보를 추출하기 위한 연구가 시도 되고 있다. 기존의 연구에서는 수집된 텍스트 데이터의 경향을 파악하기위한핵심어를 추출하는데 있어서 주로 빈도분석을 사용하고 있다. 본 연구는 담화의 핵심단어추출 방법으로 평균적인 문서 값과 비교를 통해 핵심어를 추출하는 TFIDF(Term Frequency-Inverse Document Frequency)분석법을 사용하였다는 점에서 선행연구들과 텍스트 마이닝 방법론의 차이를 가지고 있다.

본 연구에서는 문제탐색과 의사결정을 위한 빅데이터 과학의 일환인 코퍼스 언어학 기반의 내용분석수법(Corpus Linguistic Based Content Analysis)을 응용하여, 민원데이터 속에 내포되어 있는 시민들의 실질적인 공적관심사(Public Concern)를 객관적으로 기술하고 자동적으로 집약하는 공적관심사 평가수법(Public Concern Assessment Methodology)을 개발하는 것을 목표로 한다. 구체적으로 1) TFIDF(Term Frequency Inverse Document Frequency) 수법을 이용한 주요민원 분석 방법론과, 2) 공기어(Co-Occurrences) 추출과 공기어구조맵을 시각화하여 주요민원에 대한 공통 인식 및 의견대립 분석방법론을 제안하고, 문제점 및 정책적 과제를 검토하는 것으로 한다. 이를

위해 “부산시에 바란다”라는 전자민원창구에 2015년 1월 1일 - 2015년 6월 30일에 공개 접수된 민원데이터에 대한 코퍼스 분석을 실시한다.

II. 민원분석

1. 전자민원

전자민원은 민원처리의 효율화뿐만 아니라 시민참여를 통한 지역혁신의 기틀이 된다. 전자민원을 통하여 주민들은 개개인 고유의 언어체계로 자신들의 관심(concern)과 요구사항(needs)을 자유로이 표현할 수 있고, 또한 다른 사람들의 관심과 요구사항을 접하면서, 사적인 생활불편에 대한 개선요구(civil complaints) 뿐만 아니라 지역공동체의 가치를 이해하고 지역발전에 필요한 개선점을 제안할 수 있다. 그리고 공공서비스의 계획 초기단계부터 시민들의 지속적인 참여를 보증하고 제시된 의견은 공공서비스의 의사결정에 필요한 정보로서 명시화, 형식화 된다.

민원분석은 시민들의 다양한 요구사항에 대한 행정의 대응성을 향상시키고, 시민의 관점과 사회경제 정세를 고려한 객관적인 공공서비스의 평가를 통해서 미진한 점을 개선하며, 중요한 의미를 지닌 정보를 추출하고 집약함으로써 질 높은 공공서비스에 관한 의사결정의 기초자료로 활용할 수 있다. 더욱이 이해관계자간의 커뮤니케이션을 촉진하고 공공서비스에 관한 객관적인 평가 정보를 활발히 공유하며 상호간의 이해를 촉진하는 “사회적 정당화 프로세스 (Social Justification Process)”로서의 역할도 기대할 수 있다.

그러나, 민원내용분석 체계가 아직 정립되지 못하고 초보적인 단계에 머물러있다. 국민권익위원회 민원정보분석센터(2013)는 국민신문고의 전자민원을 대상으로 키워드 분석 및 성별, 지역 등 민원인의 특성을 고려한 크로스분석을 통하여 주요민원내용 요약 및 민원 경향을 분석하였다. 그러나 정부 민원데이터는 거시적 분석은 가능하나 지역별 특성을 고려한 미시적 분석이 불가능하고, 또한 현재와 같이 조사자가 미리 정해놓은 키워드로 분석하는 방법은 객관성을 보장하고 잠재되어 있는 정보의 누락이라는 한계점을 내포하고 있다.

본 연구에서는 이러한 문제점을 보완하고자 자연언어처리 수법과 내용분석 기법을 응용한 미시적 민원정보분석 방법론을 제안한다. 이를 통해 다양한 민원인의 개별적 문맥을 고려한 키워드 추출, 공통민원 탐색, 상충된 민원 탐색, 애로민원 탐색 등 정책입안자의 고충을 털어줄 수 있는 방법론을 제시할 수 있을 것으로 기대된다.

2. 수요대응성, 설명책임 및 정통성

행정은 시민들로부터 위탁 받은 공적자원을 사용하여 공공서비스를 집행운영 하는 데 있어서, 그 의사결정에 대한 설명책임(Accountability)과 정통성(Legitimacy)을 보장하여, 국민의 신뢰를 확보하

도록 요구 받고 있다. 설명책임의 전통적 의미는 대담책임(answerability)이지만, 설명책임의 궁극적인 목적은 시민의 요구와 우선사항에 대한 행정의 대응성(responsiveness)을 보장하는 것이다(Dunn, 1997). 그리고, 정치적인 책임은 공공서비스에 대한 철저한 감시로 증명되어야 한다. 전통적인 설명책임이 프로세스, 계층적 제어, 신뢰, 문화적 전통에 근거하였다면, 신공공관리(New public management)의 설명책임은 성과, 경쟁, 투명성, 계약관계에 근거하고 있다. 특히, 시민의 관점에서 공공서비스의 내용을 설명하는 것이 요구되는데, 여기서의 법률의 준수나 절차의 타당성이 아닌, 업무의 효율성과 유효성이 요구되는 “퍼포먼스 어카운터빌리티(Performance Accountability)”나 “프로그램 어카운터빌리티(Program Accountability)”에 의해 행정성적을 평가 받게 된다. 또, 행정서비스를 민간부문으로 위탁함으로써, 행정이 독점적으로 공공서비스를 제공하는 것이 아니라 시장 메커니즘을 통해서 공공서비스를 선택하는 것이 가능하게 된다.

공공서비스는 수많은 이해관계자들에게 직/간접적으로 영향을 미친다. 이러한 공공서비스의 의사결정에 대해서 이해관계자들은 서로 다른 인식을 가지고, 각자의 입장에서 서서 요구사항을 이야기한다. 그러나 모든 이해관계자들을 다 만족시킬 수 있는 공공서비스를 실행하는 것은 현실적으로 불가능하다. 여기서 “어떤 입장의 의견과 요구를 타당한 것으로 볼 것인가?”라는 정통성(legitimacy)의 문제가 대두된다. 정통성에 대해서, Jeong(2008)은 “공공서비스의 의사결정 프로세스에 있어서 정통성은 어떤 의견이나 요구사항이 해당 사회시스템에 적합하고 바람직한 것이라는 전체의 승인이 있을 때 성립되는 것으로, 승인된 의견이나 요구사항은 합리성과 공평성을 전제로 한다. 또, 그것이 민주주의적으로 정통화되지 않으면 안된다”고 설명하고 있다.

정통성의 개념은 다의적(Maurer, 1971; Pfeffer, 1981; Pfeffer, 1978; Meyer, 1983)이지만, 본 고에서는 Suchman(1995)에 따른, “어떤 주체 및 그 행위를 규범, 가치, 신념, 정의 등이 사회적으로 구조화된 시스템 속에서 바람직하고 타당하며, 혹은 적절하다고 하는 일반화된 인식”이라고 정의한다. Suchman(1995)은 실용적 정통성(pragmatic legitimacy), 도덕적 정통성(moral legitimacy), 인식적 정통성(cognitive legitimacy)이라는 3개의 정통성 개념을 제시한다(Suchman, 1995). 실용적 정통성은, 어떤 주체의 행위가 관련된 사람들의 이익 증진에 이어지는지를 따진다. 비용 편익 분석 등이 이용된다. 그러나 관련된 모든 주체가 이익을 받는 것을 보증하는 것은 실질적으로 불가능하다. 그러므로, 실용적 정통성의 개념만으로 정통성을 주장하기에는 한계가 있다. 도덕적 정통성은 행위가 옳고 그르냐는 평가에 근거한 것이며, 행위의 결과, 행위의 절차, 행위주체의 유인·보수 구조가 도덕적인지를 따진다. 인식적 정통성은 이익이나 평가가 아니라, 사회적으로 필요성이 인식되는 것에 근거한 정통성이다. 전자 민원의 코퍼스분석을 통한 사회적 이해관심의 도출은 이러한 인식적 정통성의 근거로 제시되어질 수 있다.

전자민원은 기존의 행정의 자기평가에 의한 공공서비스 계획프로세스와는 다른 것으로, 의사결정 프로세스에 있어서 사회적 커뮤니케이션을 통한 다면적인 상호평가(a wide range evaluation)와 협력적인 상호조정(corporative mutual adjustment)를 중시하는 “대화형 계획프로세스(communicative planning process)”로 특징지을 수 있다. 그러나 아직 다면적인 상호작용이 유효하게 기능하는 단계에 도달하지는 못했다. 이러한 새로운 행정서비스를 촉진하기 위해서 “대화에 의한 의견형성”과 “정

확한 정보공유"를 통한 시민의 요구사항에 대한 설명책임과 정통성 확보가 중요하다.

3. 민원 코퍼스 분석

최근, 정보처리기술을 이용해서, 사람들이 가지고 있는 지식을 체계화하고, 대규모 지식기반을 구축하여 공유하려고 하는 시도가 진행되고 있다. 정보처리기술의 기반이 되는 것이 코퍼스(Corpus)이다. 코퍼스란 텍스트집합(Text Collection)으로 사람들이 어떤 언어를 사용해왔는지를 나타내는 대규모 샘플을 의미한다. 그리고, 코퍼스언어학(Corpus Linguistics)은 대규모 텍스트집합으로부터 얻어진 관찰데이터를 주요한 증거로 활용하고, 언어나 구의 용법 및 의미를 설명하려고 하는 언어연구의 하나의 방법론이다. 그 특징은 실제적으로 사용된 언어데이터를 분석대상으로, 컴퓨터를 이용한 재현가능한 계산수법이라는 점이며, 이러한 점에 있어서 코퍼스 언어학은 경험주의적인 관찰수법이라고 말한다. 코퍼스의 이용은 정보검색, 문서관리, 학습지원 등 다양하지만, 민원에 대한 행정의 수요대응성을 향상시키기 위해서도 이용할 수 있다. 전자민원의 발언은 물론, 미디어나 전문학술자료 등의 다양한 분야의 정보가 코퍼스로 축적 되어있다. 이러한 민원관련 코퍼스를 행정서비스에 필요한 정보인프라로 공유하고 재이용 함으로써, 사회적 요청을 고려한 효율적인 행정서비스를 실현할 수 있을 것으로 기대된다. 그리고 민원정보인프라는 행정의 의사결정내용과 그 프로세스를 빠르게 전달하는 것을 보증한다. 또 코퍼스를 이용한 정보인프라구축에는 계획적 또는 지속적으로 질 높은 판단정보를 유지관리하기 위한 민원코퍼스 인프라기반을 구축할 필요가 있다.

전자민원 코퍼스를 이용하여 시민들의 공적관심사를 집약하고 평가하기 위해서, 반드시 갖추어야 할 2가지 조건이 있는데 민원내용분석의 객관성과 정확성이 그것이다. 즉, 첫째, 주민이나 기업 등 다양한 이해관계자들의 발언 속에 포함되어 있는 관심사를 객관적으로 알기 쉽게 기술하는 것, 둘째, 편견 없이 올바르게 관심사를 평가하고 분류하기 위한 적합한 수법을 개발하는 것이 필요하다. 지금까지 시민들의 공적관심사에 대해서 다양한 내용분석이 시도되어 왔으나, 그 대부분이 조사자가 설계한 설문조사를 통해서 공적관심사를 집약하는 방법이었다. 이러한 기존의 방법은 조사자의 주관에 의존하기 때문에 조사자가 달라질 경우, 분석결과가 판이하게 달라질 수 있으며 또 분석결과를 자의적으로 조작할 수 있는 문제점이 지적되고 있다. 이에 비해, 본 연구에서 제안하는 전자민원 코퍼스를 이용한 공적관심사 평가방법론은 코퍼스 언어학을 기반으로 한 계산론적 접근법을 응용한 민원내용분석법으로 다음 3가지 장점이 있다. 첫째, 민원인의 발언을 있는 그대로 분석하기 때문에 조사자가 미리 예상하지 못했던 새로운 공적관심사를 발굴할 수 있다. 둘째, 서로 다른 조사자가 분석해도 같은 결과를 도출하는 재현가능성을 보장하여 분석결과와 객관성을 충분히 확보할 수 있다. 셋째, 적은 비용과 인력으로 방대한 민원데이터를 용이하게 분석할 수 있다. 이에 따라, 공공서비스의 의사결정에 민원을 근거자료로 사용하는 것에 대한 신뢰성을 확립할 수 있다. 즉, 시민들은 자신이 관심을 가지고 있는 공공서비스에 대해서 다른 사회구성원들이 어떻게 생각하고 있는지 쉽게 검색할 수 있고, 행정은 공공서비스에 대한 의사결정에 있어서 필요한

근거자료로 집약된 공적관심사를 사용할 수 있다.

Ⅲ. 민원코퍼스분석 방법론

1. TFIDF

본 연구에서는 GDA메타텍스트(GDA Meta Text)를 기초로 하여 민원의 주요키워드를 추출하기 위해 TFIDF(Term Frequency Inverse Document Frequency)분석을 실시한다. TFIDF법은 문서중의 단어수의 중요도를 판정하여 중요하다고 판정되는 언어만 분석의 요소로 사용하는 방법이다 (Salton and McGill, 1983).

TFIDF w,a 는 중요도의 총량을 나타내는 것으로, TFIDF스코어 라고 불리며, “문서 a에 나타나는 단어w의 빈도인 $TF_{w,a}$ ”와 “비교대상이 되는 문서 중 w가 포함된 문서의 수 IDF_w ”를 곱하여 얻을 수 있다. TFIDF스코어가 높다는 것은 어떤 문서 a에서 단어 w가 중요한 의미를 가지는 것을 의미한다. 어떤 단어가 한 문서 안에서 자주 나타나면 높은 단어빈도수를 가지지만, 그렇다고 이 단어가 반드시 중요한 것은 아니다. 해당 단어가 중요한지 아닌지는 다른 문서와의 비교를 통해서만 가능하다. 즉 비교문서의 빈도가 중요한 것이다. IDF_w 의 값은 전체 문서값 DF_w 의 로그지수로 정의한다. 즉, 얼마나 많은 문서에 단어 w가 출현하는지를 나타내는 것이다. 예를 들면, 만약 단어 w_1 이 몇 몇 되지 않는 문서에 출현하고, 분석용 문서에서는 자주 출현한다면, w_1 의 IDF 값은 높고, TFIDF스코어도 높아진다. 이와는 반대로, 어떤 단어 w_2 가 많은 문서에서 출현하지만, 분석용 문서에는 적게 출현한다면, 단어 w_2 의 IDF값은 낮아지고, TFIDF스코어 또한 낮아진다.

〈식 1〉 TFIDF값

$$TFIDF_{w,a} = TF_{w,a} * IDF_w \quad \langle \text{식 1} \rangle$$

$$IDF_w = \log(N / DF_w) + 1$$

$TF_{w,a}$ = Number of occurrences of term w in document a
 DF_w = Number of documents containing term w
 N = Total number of documents

〈식 1〉에서 나타내는 것처럼 IDF값과 TF값이 높으면 높을수록, TF-IDF값이 높다. j문서의 i단어의 수를 예로 들면 TF-IDF값이 높은 i단어는 j문서에서 중요한 단어이며, 토픽이라고 할 수 있다. 본 연구의 DF(Document Frequency) 비교문헌 값으로는 카이스트가 제공하는 620건의 신문데이터를 이용하였다.

2. 공기어구조맵

주요민원키워드에 대한 민원인들의 인식을 핵심단어 간의 구조적인 형태로 확인하기 위하여, 공기어구조맵(Co-Occurrences Structure Map)을 작성한다. 우선, 민원발언으로부터 키워드에 관한 공기어와 그 공기빈도를 도출한다. 여기서는 앞서 TFIDF수법을 이용하여 추출한 키워드(TFIDF치가 높은 단어)를 중심어로 지정하고, 중심어가 포함된 민원발언의 전후 한 문장내의 단어를 공기어로 검색한다. 또한 이 수법을 발전시켜, 공기어 벡터를 구하여 민원인들 간의 인식의 차이를 계산할 수 있다. 이것은 공기어 벡터들 사이의 코싸인(Cosine) 거리를 계측함으로써 가능하다. 공기어 추출로부터 공기어 벡터의 유의도를 구하기 위한 프로그램으로 벡터 공간법 VSM(Vector Space Machine)를 이용할 수 있다.

민원인의 발언내용의 차이를 표현하는 척도로, 공기데이터로부터 벡터 공간법 VSM(Vector Space Machine)을 이용하여 민원인 간의 발화 유의도를 구한다. VSM수법은 문서를 다차원공간에 있어서 벡터로 표현하는 정보검색 (Information Retrieval, IR)모델의 하나로, 텍스트의 내용을 텍스트내의 단어를 기초로 한 특성 벡터로 표현하는 방법이다. VSM수법에 의해서, 하나의 텍스트가 하나의 벡터로 표현되고, 서로 다른 텍스트간의 유의도가 벡터간의 유의도로 구해진다. 본 연구에서는 텍스트간의 문서비교가 아닌, 주요민원에 대한 인식의 차이가 있는지 확인하기 위해서, 각각의 키워드에 대해서 공기어 벡터를 구하고 이때, 각 민원하나를 벡터로 표현한다. 그런 다음 민원사이의 유의도를 계산한다.

구체적으로, 어떤 키워드 k 와 공기하는 모든 단어 $w=\{w_1, \dots, w_n\}$ 은, 각각의 공기빈도 $TFw=\{TFw_1, k, \dots, TFw_n, k\}$ 로 그 가중치를 부여한다. 여기서 n 은 하나의 텍스트안에 있는 모든 단어의 수를 나타낸다. 그리고 n 차원에 있어서 각각의 민원 p 의 공기데이터의 벡터 $\rightarrow TFw, t, p$ 를 아래와 같이 표현할 수 있다.

$$\rightarrow TFw=\{TFw_{1,t,p} \dots TFw_{n,t,p}\} = R_n$$

본 연구에서는 민원인을 특정하지 않고, 민원전체에 대한 공기어구조를 살펴보았다.

IV. 사례분석: 부산시 민원

1. 주요민원

“부산시에 바란다”라는 전자민원창구에 2015년 1월 1일 - 2015년 6월 31일에 공개 접수된 1336건의 민원데이터에 대한 코퍼스분석을 실시하였다. 우선 주요민원을 확인하기 위하여 TFIDF분석을 실시하였다. <표 1>은 TFIDF값에 따라 상위 10개의 단어를 집계한 것이다.

2015년 상반기 전체 민원 중 출현빈도가 높은 상위 10개의 단어는 버스, 식품, 노선, 시민, 설치, 안전, 공무원, 연금, 아파트, 차량이었다. 주요키워드의 출현특성을 확인하기 위하여 월별출현분포분석을 실시하였다. 상위10위의 키워드로 출현분포가 유사한 쌍으로 버스-노선, 식품-안전, 공무원-연금이 확인되었다. 이는 2015년 상반기에는 버스노선, 식품안전, 공무원연금이 주요민원으로 제기된 것을 알 수 있다.

〈표 1〉 부산시 민원 TFIDF분석

순위	단어	TFIDF값(소수점 이하 생략)
1	버스	3658
2	식품	2747
3	노선	2145
4	시민	1835
5	설치	1817
6	안전	1651
7	공무원	1534
8	연금	1489
9	아파트	1247
10	차량	1170

다음으로, 월별로 TFIDF분석을 실시하여 6개월간 주요민원의 동향을 살펴보았다. 〈표 2〉는 그 결과를 정리한 것이다. 2015년도 상반기민원은 정관 신도시 아파트 입주자들의 민원이 주를 이루고 있다. 정관신도시에서 울산으로 가는 버스노선 및 기타 버스관련 민원이 지속적으로 제기되고 있는 것을 알 수 있다. 1월에는 정류장, 회전, 단속, 신호, 시간, 도로가 특징적으로 언급되고 있다. 버스정류장에 대한 민원과 도로사정에 관한 민원이 주를 이루고 있다. 2월에는 답변, 주민, 분양, 민원, 도서관, 아동이 다른 달과 비교하여 특징적으로 언급되고 있다. 민원에 대한 답변에 대한 문제제기가 일어나고 있음을 알 수 있다.

〈표 2〉 2015상반기 부산시 민원 월별 TFIDF분석

순위	1월		2월		3월		4월		5월		6월	
1	버스	973	버스	404	버스	448	버스	545	식품	2073	병원	880
2	노선	492	식품	290	노선	392	화소	304	연금	1438	버스	769
3	차량	377	아파트	269	설치	279	노선	288	공무원	1114	환자	745
4	정류장	275	답변	242	정관	201	정관	267	안전	981	노선	675
5	아파트	262	주민	208	아파트	193	설치	213	제안자	847	동선	587
6	회전	260	분양	194	시공사	193	공사	200	무상	774	설치	523
7	단속	257	민원	192	입주	186	호기	196	버스	759	공개	519
8	신호	252	도서관	190	주차장	186	대중교통	186	보육	706	화소	456
9	시간	235	차량	186	주차	169	주민	185	급식	565	격리	431
10	도로	225	아동	182	차량	160	고리	181	기초	524	확진자	401

3월에는 시공사, 입주, 주차장, 주차가 특징적으로 언급되고 있으며, 주차문제에 관한 민원이 주로 나타나고 있음을 알 수 있다. 4월에는 화소, 공사, 호기, 대중교통, 고리가 특징적으로 나타나고 있다. 고리와 호기는 고리 원전에 관한 민원이며, 화소는 고화소 CCTV설치에 관한 민원이다. 5월에는 공무원 연금과 기초연금에 관한 언급과 무상급식에 관한 민원이 주를 이루고 있다. 6월에는 메르스 사태와 관련한 단어들인 주로 언급되었다. 병원, 환자, 동선, 공개, 격리, 확진자가 이와 관련된 단어로 6월 상위 토픽의 주를 이루고 있다. 또한 구체적인 지명으로는 정관과 고리가 언급되고 있다.

정리하면, 상위 10개의 토픽으로부터 시민들의 민원에서 주로 나타나고 있는 구체적인 문제로 버스노선, 정관신도시의 기반시설 미비, 고화질 CCTV설치요구, 공무원연금과 기초연금, 무상급식, 메르스 대응, 고리원전가동, 식품안전에 관한 불안 등이 있다는 사실을 알 수 있다. 월별로 나타나는 토픽의 변화를 전체 민원과의 비교를 통해 선명히 할 수 있으며 맥락을 벗어나서 직관적으로 파악하는 것이 가능하다는 사실을 알 수 있다.

2. 주요민원에 대한 공기어구조맵

주요민원 키워드에 대해서 구체적으로 어떤 내용들이 언급되었는지 알기 쉽게 가시화하기 위하여 공기어구조맵(Co-Occurrences Structure Map)을 작성하였다. 여기서는, 2015년 1월분 민원에 대해서 TFIDF값이 가장 높은 “버스노선”을 주요민원 키워드로 하고, “버스노선”에 대한 주요공기어 (공기어 출현빈도 10회이상)를 추출하여, 공기어구조맵을 작성하였다. 즉, 키워드 A (“버스노선”)과 함께 출현하는 임의의 단어 B를 동시에 내포하는 민원수를 측정하여 10회가 넘으면, B를 A의 주요 공기어로 정의한다. 공기어구조맵 작성에는 Graphviz프로그램을 이용하였다.

다음으로, <표 3>을 바탕으로 공기어네트워크구조를 작성하여, 주요키워드에 대한 이해관심사항이 무엇인지 이해하기 쉽도록 시각화하였다. 그림1의 예에서는, 제1층 주요공기어 “차량”, “시민”, “아파트”, “설치”들과 공통으로 출현하는 단어로 제2층 주요 공기어에 “과속”, “시간”, “안전”, “불편” 세 단어가 있다. 이 세 단어가 민원인들의 버스노선에 대한 주요이해관심사(concern)라고 해석 할 수 있다.

또한 제1층 주요공기어 “차량”, “시민”, “아파트”, “설치”와 각각 연결된 단어들을 통하여, 각각의 부주제에 대한 이해관심구조를 이해하는 데 도움이 된다. 예를 들면, “차량”에 대해서는 “올림픽경기장”, “우회전”, “차선”, “공사” 등 특정지역 및 특정 인프라에 대한 관심이 집중되어 있는 것을 알 수 있다. 즉, 민원내용을 용이하게 요약하고 그 특성을 정리할 수 있다.

〈표 3〉 “버스노선”의 공기어

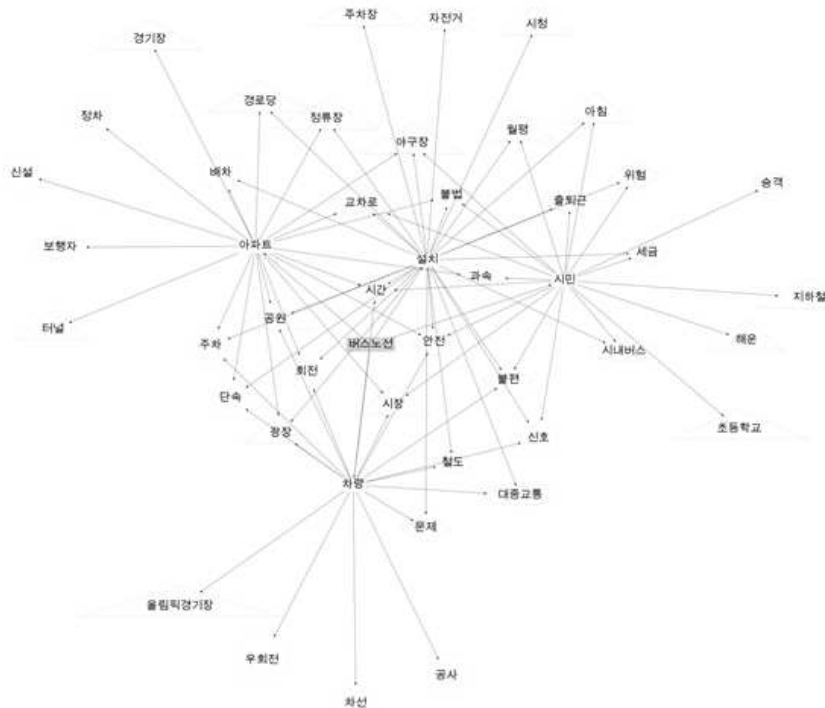
제1차 공기어	차량		시민		설치		아파트	
제2차 공기어	시간 시장 단속 문제 신호 안전 회전 불편 철도	공원 주차 공사 차선 올림픽경기장 광장 우회전 대중교통	시간 시장 신호 안전 불편 시내버스 출퇴근 야구장 아침 해운	위험 교차로 승객 지하철 세금 불법 월평 초등학교 과속	시간 단속 문제 신호 안전 불편 회전 철도 공원 주차 시내버스 야구장 출퇴근 배차	아침 위험 정류장 광장 대중교통 세금 주차장 불법 시청 월평 경로당 교차로 자전거	시간 시장 단속 안전 회전 정류장 공원 주차 정차 경기장	야구장 터널 배차 광장 불법 보행자 경로당 교차로 신설 과속

*제1차 공기어: “버스노선”에 대한 주요공기어 (공기어 출현빈도 10회이상)

*제2차 공기어: 제1차 공기어에 대한 주요공기어 (공기어 출현빈도 10회이상)

또한 이러한 공기어구조맵을 시간별로 요약함으로써 민원의 주제에 따른 패턴특성 분석이 가능하다. 이러한 분석을 통하여 민원의 특성 (단발성/지속성, 특정집단/포괄적집단 등)에 따른 대응 메뉴얼 작성, 민원수요예측, (책임설명 및 투명성 확보를 위한) 민원처리의 정보공개 등에 실무적인 도움을 줄 수 있을 것으로 기대된다. 또한 위 분석결과는 민원통계백서 등으로 활용할 수 있을 것이다.

〈그림 1〉 “버스노선”에 대한 공기어구조맵 (2015년 1월분)



V. 결론

본 연구에서는 방대한 민원데이터의 내용을 객관적이고 용이하게 요약하기 위하여 코퍼스분석 기법을 응용하여 공기어구조맵을 작성하는 방법론을 제안하였다.

“부산시에 바란다”라는 전자민원창구에 2015년 1월 1일 - 2015년 6월 31일에 공개 접수된 민원 1336건의 데이터에 대한 코퍼스분석을 통해 시민들의 민원에서 주로 나타나고 있는 구체적인 문제로 버스노선, 정관신도시의 기반시설 미비, 고화질 CCTV설치요구, 공무원연금과 기초연금, 무상급식, 메르스 대응, 고리원전가동, 식품안전에 관한 불안 등이 있다는 사실을 알 수 있었다. 월별로 나타나는 토픽의 변화를 선명히 할 수 있으며, 핵심어 파악을 통해 대량의 민원에 내포된 공적 관심사를 맥락을 벗어나서 직관적으로 파악하는 것이 가능하였다. 또한, 2015년 1월분 민원에 대해서 TFIDF값이 가장 높은 “버스노선”을 키워드로 공기어구조맵을 작성하여 버스노선과 관련된 시민들의 이해관심을 맥락으로 파악할 수 있다. 버스노선의 공기어인 “차량”, “시민”, “아파트”, “설치”와 연결된 단어들을 통하여, 각각의 부주제에 대한 이해관심구조를 이해할 수 있었다. 구체적으로 “차량”에 대해서는 “올림픽경기장”, “우회전”, “차선”, “공사” 등 특정지역 및 특정 인프라에 대한 관심이 집중되어 있는 것을 알 수 있었다.

이러한 통계적기법을 응용함으로써 체계적/효율적인 민원데이터 관리를 꾀할 수 있으며, 민원 대응평가 및 정보공개시스템과 연계하여 민원 관련 자료를 통합함으로써 종합적인 민원정보시스템을 개발할 수 있다. 또한 이를 통하여 사회적으로 필요성이 인식되는 것에 근거한 인식적 정통성의 근거가 되어 민원행정에 대한 설명책임 및 정통성 확보에 기여할 수 있을 것으로 기대된다.

특히, 주요민원 키워드 및 공기어구조맵의 시계열분석은 주요키워드에 대한 민원키워드에 대한 패턴특성분석 뿐만 아니라, 공적이해관심사(Public Concern), 인식체계 (Cognitive System), 인식의 변화(Perception Change)를 이해하는 데 도움이 된다. 이러한 분석을 통하여 민원의 특성 (단발성/지속성, 특정집단/포괄적집단 등)에 따른 대응메뉴얼 작성, 민원수요예측, 민원처리의 정보공개 등 실무적으로도 도움을 줄 수 있을 것으로 기대된다.

본 연구는 민원데이터의 객관적이고 효율적인 요약을 통하여 시민들의 이해관심을 집약하여 파악하는 것을 목적으로 하였다. 하지만 한발 더 나아가 이를 통하여 미래의 정책수요를 예측하고 파악하기 위한 연구의 전개가 요구된다. 이를 위해서는 민원코퍼스의 주기적인 분석 시스템 구축과 코퍼스에 대한 데이터베이스화 및 검색시스템 구축을 통하여 장기적인 추세와 경향을 파악하는 연구가 필요하다.

참고문헌

이태현·홍순구·정하영·홍혜승. (2017). 지방정부 공공사업에 관한 일간지 미디어자료의 시계열

- 코퍼스분석: 기장 해수담수화플랜트사업을 중심으로. 「지방정부연구」, 20(4): 1-23.
- Motomu, N.,(2006). An Introduction to Topic Maps, TDU
- Dunn, D. Politics and Administration at the Top: Lessons from Down Under, Pittsburgh University of Pittsburgh Press.
- Jeong, H.Y,(2008). A Study on Public Deliberation in Provision of Social Overhead Capital, Kyoto University (in Japanese)
- Hatori, T., Kobayashi, K., and Jeong, H., (2013). Discourse Theory and Normative Evaluation of Public Debates. Journal of Japan Society of Civil Engineers, Ser. D3 (Infrastructure Planning and Management), 69(2): 101-120 (in Japanese)
- Jeong, H., Kobayashi, K., Hatori, K., and Shiramatsu, S.,(2010). Facet Decomposition and Discourse Analysis of Public Debate. Journal of Japan Society of Civil Engineers, Ser. F (Construction and Management), 66(1): 45-56 (in Japanese)
- Jeong, H., Shiramatsu, S., Kobayashi, K., and Hatori, T.,(2008). Discourse Analysis of Public Debates Using Corpus Linguistic Methodologies. Journal of Computers, 3(8): 58-68
- Jeong, H., Hatori, T. and Kobayashi, K.,(2007). Discourse Analysis of Public Debates: a Corpus-Based Approach. Proceedings of the IEEE Systems, Man, and Cybernetics Conference, Montréal, Canada, 7-10 October: 1782-1793
- Pfeffer, J.(1981). Management as symbolic action: the creation and maintenance of organizational paradigms, In: Cummings, L.L. and Staw, B.M. (eds): Research in Organizational Behavior, Vol.13: 1-52, JAI Press
- Pfeffer, J., Salancik, G.(1978). The External Control of Organizations: A Resource Dependence Perspective, Harper and Row
- Meyer, J.W., Scott, W.R.(1983). Centralization and the legitimacy problems of local government, In: Meyer, J.W. and Scott, W.R. (eds): Organizational Environments: Ritual and Rationality, : 199-215, Sage
- Suchman, M.C.(1995). Managing legitimacy: strategic and institutional approaches, Academy of Management Review, 20(3): 571-610

정하영(鄭蝦榮): 동아대학교 도시공학 학사 및 석사, 일본교토대학 도시사회공학연구과 박사를 취득하였으며, 일본교토대학 도시사회공학연구과 인간안전보장공학연구 프로젝트의 연구원 및 경영대학원 아시아 비즈니스 리더육성강좌의 조교수를 거쳐, 현재 동아대학교 경영문제연구소 공동연구원 및 주식회사 만타의 대표이사를 겸임하고 있다. 주요 연구로는 통계학, 계량경제학, 정보학, 자연언어처리학을 기반으로 한 '삶의 질 지표 개발', '주민참여형 공공계획의 코퍼스분석기법 및 온톨로지 개발', '트위터 내용분석을 통한 불안계량기법 개발', '지역아이덴티티의 정량적 분석기법 개발' 등 있다.

이태헌(李泰憲): 동아대학교에서 도시계획을 전공하고 일본의 리츠메이칸(立命館)대학 정책과학부에서 정책과학 박사학위를 취득하였다. 주요 연구로는 일본 니시진 지역의 토지이용변화와 마을만들기에 관한 연구, 사상구 토지이용계획 정책에 관한연구가 있다. 현재는 동아대학교 산학협력단 전임연구원으로 재직하고 있으며, 주요관심분야는 마을만들기를 통한 부산의 지역혁신, Co-creation, 지리정보시스템을 활용한 지역정책 분석 및 새로운 정책대안 모색이다.

홍순구(洪淳玖): University of Nebraska-Lincoln에서 경영정보학 박사학위를 취득하고 동아대학교 경영정보학과 교수로 재직중이다. 학문적 관심분야는 Co-creation, 웹 접근성, 정보시스템 평가, 지역혁신 등이고 주요저술로는 「유비쿼터스의 이해와 활용사례」, 「정보기술의 전략적 활용: 경영정보시스템」, 주요논문으로는 “Co-creation 플랫폼 운영 전략: 삼성 앱스와 애플 앱스토어 비교(2014)”, “ICT기반 지역공동체 가치창출 전략의 탐색”(2013), “청년층 일자리 창출문제에서 Co-creation적 해결방안에 관한 연구: 부산광역시를 중심으로”(2015) 등이 있다. 주요경력은 Texas A&M International University 조교수, 한국은행 근무, 동아대학교 재무처장, 기획처장 등이다(shong@dau.ac.kr).

Abstract**A Copus Analysis of Electronic Petitions For Improving the Responsiveness of Public Services: Focusing on Busan Petition**

Jeong, Hayeong

Lee, Tai Hun

Hong, Soon-Goo

So far, there are diverse content analyses on public concerns that have been attempted. However, most of them have been conducted based on a survey questionnaire designed by the researcher. The survey questionnaire highly depends on the researchers' subjective point of view. If the researchers are different, the results of the analysis may also differ. It requires great effort for the researcher to collect the data. Regarding the problem on high dependency on the researcher, this study proposes a corpus-linguistics-based content analysis as a part of big data science to search latent problem and to support decision-making, and it also attempts to develop a public concern assessment methodology to describe objectively the actual public concerns from civil petitions data and automatically summarizes them. For this purpose, this study proposed two methodologies - i) new content analysis of civil petitions to apply TFIDF (Term Frequency Inverse Document Frequency) methodology and ii) Co-occurrences extraction and co-occurrences structure map to visualize a common perspective on a policy issue and any conflict of public concerns. As a case study, the proposed content analysis is conducted with open civil petitions that were submitted to Busan electronic civil petition window, named "Wish to Busan (Busansi-e-baranda)" from January 1, 2015 to June 30, 2015. The results can be utilized as a basic evidence for policy making such as key concern of civil petition, common perspective, and conflict of opinion.

Key Words: Regional Innovation, Busan electronic civil petition, Corpus Analysis, Text Mining, Public Concerns