「지방정부연구」 제20권 제4호(2017 겨울): 1-23
The Korean Journal of Local Government Studies, Vol.20 No.4 (2017 Winter)
http://dx.doi.org/10.20484/klog.20.4.1

지방정부 공공사업에 관한 일간지 미디어자료의 시계열 코퍼스분석: 기장 해수담수화플랜트사업을 중심으로*

이 태 헌**
홍 순 구***
정 하 영****
홍 혜 승*****

국문요약

본 연구는 이해관계자간의 인식차이를 객관화하기 위하여 시계열로 핵심어를 추출하는 과정을 통해 대립이나 갈등을 해소하기 위한 단서를 제공할 수 있는 자동화된 텍스트분석 방법론을 구축하는 것을 목적으로 한다. 이를 위해 현재 논쟁이 진행되고 있는 해수담수화 플랜트사업에 관한 거시적인 관점에서 문제인식의 차이를 이해하기 위해 지역 2개 신문사와 중앙 3개 신문사의 기사를 수집하여 시계열 코퍼스(corpus) 분석을 실시하였다. 분석결과 해수담수화 논의에 있어서 미디어간 인식의 차이가 존재하는 것을 알 수 있었다. 본 연구는 코퍼스 분석에 있어서 한글자연어처리 기술과 TFIDF분석을 통한 핵심단어의 추출을 통해 자동적인 텍스트분석 시스템을 구축하기 위한 토대가 된다는 점에서 학문적인 기여를 하고 있다.

주제어: 지역혁신, 해수담수화플랜트, 코퍼스 분석, 텍스트 마이닝, 공동가치창출

Ⅰ. 연구의 배경 및 목적

공공사업에는 다양한 이해관계자(Stakeholder)가 관여한다. 이해관계자는 직, 간접적으로 이해관계를 가지고 영향을 주고받는 사람을 가리키는 넓은 의미로 해석되며, 공공사업의 경우 일반시민이 이에 해당된다. 실제로 시민은 미디어를 통해서 공공사업에 관한 내용을 공유하고, 새롭게 생성되는 정보를 전파하며, 어느 이해관계에 속한 주장을 전달하는 등 사회적인 의견수렴 역할을

* 이 논문은 2015년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2015S1A3A2046781).
** 제1저자
*** 교신저자
**** 공동저자
***** 공동저자

담당하고 있다.

본 연구는 이러한 미디어를 통한 다수의 인식전달 과정을 거시적인 토론이라는 관점으로 보았다. 각기 다른 입장을 가지고 있는 다양한 이해관계자들 사이에는 문제인식의 차이, 정보량과 질의 차이, 사실관계를 해석하는 방법의 차이가 존재한다. 예를 들면 수자원 부족 문제에 대해 생활용수 및 산업용수가 부족하므로 보충해야 한다는 인식을 가지는 집단이 있는 반면, 용수 공급에 전혀 문제가 발생하지 않을 것이라는 전망을 가진 집단이 존재한다. 이러한 문제 인식의 차이는 정보의 양과 질의 차이 및 정보에 대한 해석의 차이에서 발생한다. 정보의 정확성에 대한 질적 차이와 정보량의 차이는 논점에 대한 대화가 활발히 일어나게 되면 쉽게 합의점을 도출 할 수 있다. 그러나 이해관계자의 해석의 차이는 이해관계자의 입장이라는 범위 안에서 이루어지므로 쉽게 좁혀지지 않는다. 이러한 해석의 차이는 상대방의 입장에 얼마만큼 다가갈 수 있는가에 따라서 합의점에 도달할 수 있다.

지금까지는 공공사업과 같은 사회적인 합의가 필요한 사안에 대해서 전문위원회를 통해 전문가의 검토를 거치거나 주민대표의 참가를 통한 미시적인 토의절차를 거쳐 사회적 합의의 근거를 마련해 왔다. 하지만 거시적인 담화를 분석할 수 있는 기술이 발전하게 된다면 이를 토대로 사회적 합의의 근거를 더욱 객관적으로 도출할 수 있을 것이다.

거시적인 토론에서 상반된 견해를 가진 집단들이 상대방의 입장으로 다가가기 위해서는 해결되어야 할 문제가 여러 가지가 있지만, 그 중에서 중요한 하나는 입장 차이가 있더라도 공통된 하나의 주제 혹은 핵심어를 가지고 대화를 진행하는 노력이다. 하지만 논쟁이 심화되는 공공사업의 경우에는 대부분 이해관계자 입장에서 가까운 주제(핵심어)를 중심으로 토론이 일어나게 되므로 합의점 도출에 접근하지 못하고 제자리걸음을 하게 된다.

미시적인 토론 과정에서는 주고받는 대화를 통해 이러한 차이를 수정해가는 과정과 상대방의 입장이 수정되었음을 확인해가는 인식과정이 존재한다. 또한 담화를 통해 하나의 주제에서 다음 주제로 진행되어 가는 과정을 우리는 회의기록상의 맥락이 있는 서술형태문장을 읽으면서 파악이 가능하다. 그러나 거시적인 토론에서는 방대한 양의 비정형 데이터로부터 각기 분산된 이해관계자의 핵심어를 추출하는 것이 어려우며 미시적 토론에서 가능한 맥락에 따른 핵심어의 진행과정을 파악하는 일이 쉽지 않다.

거시적인 토론을 통해 이해당사자간 의견을 좁혀가기 위해서는 각기 다른 입장에서 다른 주제(핵심어)를 중심으로 발생하는 문제해석의 차이를 좁히는 과정은 반드시 필요하다. 따라서 자동화된 텍스트 분석을 통해 이해관계자의 핵심어를 추출해내는 분석방법에 대한 연구가 요구된다.

본 연구는 이해관계자간의 인식차이를 객관화하기 위하여 시계열로 핵심어를 추출하는 과정을 통해 대립이나 갈등을 해소할 수 있는 단서를 제공할 수 있는 자동화된 텍스트분석 방법론을 구축하는 것을 목적으로 한다.

분석대상으로 이해관계자 간 첨예하게 대립하고 있는 “부산 해수담수화플랜트사업”을 선정하였다. 해수담수화 플랜트사업은 물 부족 문제를 해결하기 위해 바닷물을 담수화하는 혁신적인 기술이나 해수담수화 플랜트가 고리원전에 가까운 기장에 건설됨으로써 해당 주민의 반대여론이 확

산되고 있다. 이러한 해수담수화 플랜트사업에 관한 거시적인 토론에서 문제인식의 차이를 이해하기 위해 지역 2개 신문사와 중앙 3개 신문사의 기사를 수집하였다. 각 신문사의 해수담수화 플랜트 건설에 대한 비정형 텍스트 데이터로부터 기계적인 분석방법을 통해 인식과 의견을 단순화하여 나타내는 핵심어를 추출하는 코퍼스(corpus) 분석을 실시하였다.

논리적인 논증방법은 인과관계에 의한 연역적인 논증과 사례의 축적에 의한 귀납적인 논증 두 가지로 이루어진다. 이중에서 귀납적인 논증은 샘플링과 조사표를 통한 설문조사, 전화조사 등의 방법을 주로 사용하고 있다. 이러한 기존의 조사 및 분석방법은 다수의 데이터를 획득하기 위한 시간과 노력이 많이 들고 분석에 있어서도 조사표 회수, 타이핑, 코딩 등의 여러 가지 전 처리 절차를 거쳐야하는 어려움이 있다. 이러한 기존의 조사 및 분석 방법이 예를 들어 조사자가 일일이 셀 수 있고 처리할 수 있는 단위의 조사데이터를 토대로 분석을 실시한다고 가정했을 때 본 연구에서 제시하는 코퍼스 분석 방법론은 데이터가 존재한다는 전제 하에서 사람이 셀 수 없는 데이터의 양이 되어도 분석이 가능하다. 이러한 압도적인 양의 데이터를 통한 논증은 많은 이해관계자가 참여하는 거시적인 토론에서는 사회적 합의의 근거를 마련하기 위한 장치로 유용하게 사용될 수 있다.

Ⅱ. 이론적 배경 및 선행연구

1. 오피니언 마이닝

본 연구에서는 공공사업에 관한 거시적 토론의 분석을 위해 코퍼스 분석방법론을 제시하고 있다. Jeong et al.(2008)의 연구를 통해서, 다수가 참여하는 토론의 내용과 구조를 객관적으로 분석하기 위해서 TFIDF(Term Frequency-Inverse Document Frequency), T-VSM(Term-based Vector Space Machine) 및 MDS(Multi-Dimensional Scaling)를 포함한 코퍼스기반의 담화분석 방법론이 제시되어왔다. Jeong et al.(2008)의 연구에서는 코퍼스 분석 방법론을 적용하여 소규모 회의의 토론내용을 분석하였다. 공공사업의 사회적 적절성을 판단하기 위한 방법으로 주로 사업주체간의 회의, 주민협의회 회의, NGO단체와 지역행정과의 회의 등 작은 규모의 토론내용을 계량화하는 연구를 진행하였다.

본 연구에서는 좀 더 넓은 범위라 할 수 있는 공공사업에 대한 거시적인 사회적 담화를 분석하기 위한 연구로 확장하였다. 이를 위해 사회적 담화에 영향을 끼치는 요인으로 미디어 간 인식의 불일치 정도를 분석하고 시간에 따른 변화를 파악하는 방법을 통해 거시적인 담화의 일부분을 포착하고자 하였다. 이와 같은 연구는 빅데이터 분석을 통한 오피니언 마이닝 분야에서 주로 다루어지고 있다. 최근의 연구동향에서 오피니언 마이닝은 단순히 상품이나 서비스의 품질에 대한 대중의 의견을 분석하는 것 뿐 만 아니라, 거시적으로 다루어져야 할 사회적인 사건 및 현안에 대한 논쟁에 대해서 까지 비정형 텍스트 분석기법을 통해 분석, 종합하고 있다. 본 연구는 이러한 오피니

언 마이닝의 맥락에서 대중의 의견을 형성하는데 영향을 주는 원인요소로써의 미디어간의 인식차이와 시간에 따른 핵심인식의 변화를 분석하고자 하였다.

오피니언 마이닝에 대한 연구는 불특정 다수의 분산된 의견으로 이루어진 비정형 데이터로부터 평판분석을 통한 마케팅 분야의 적용에서 활발히 진행되어 왔으며(서지훈, 2014), 뉴스 컨텐츠 분석을 통해 주가예측의 중요정보로 활용하는 지능형투자 의사결정 모형에 관한 연구(김동영, 2014), 뉴스 투자주식의 가격지수를 이용한 미래예측(Kim et al, 2013), 온라인 쇼핑몰 상품평 검색효율을 높이기 위한 연구(Yang, 2009; Yune, 2010), 트위터 데이터를 이용하여 국내대학의 인식 및 선호도 분석(양민혁 외 3, 2014), 2012년 한국대선에 관한 분석(배정환 외 2, 2013) 등의 연구가 진행되고 있다.

대량의 비정형텍스트 데이터를 전제로 하는 텍스트 마이닝 기법은 크게 데이터 수집기술과 자연어 처리기술을 토대로 이루어 진다. 비정형데이터는 문자 그대로 고정된 필드에 저장되어 있지 않고, 온오프라인 상에 분산된 장소에 각기 다양한 형태로 존재하는 정형화 되지 않은 데이터를 뜻한다. 이러한 비정형데이터를 분석하기 위해서는 수집하는 방법에 대한 기술이 중요하며 데이터 수집기술은 웹 크롤링(Web Scraping)과 같은 연구주제에서 다루어지고 있다. 본 연구는 이미 수집된 데이터에 대한 분석방법인 자연어 처리기술을 바탕으로 진행되었다.

비정형 텍스트데이터의 분석을 위해서는 형태소 분석이 요구되므로 형태소 분석에 대한 연구가 진행 중이다. 한글을 전 처리하기 위한 형태소 분석기는 “꼬꼬마 분석기”, 자바 한국어 형태소 분석기 ”KOMORAN” 등 다수 존재하고 있다. 본 연구에서는 명사와 조사 및 어간 어미를 구분하는 조금 더 세부적인 형태소별 주석달기를 실시하기 위하여 교토대 정보연구과에서 만들어진 오픈소스 형태소 분석기 MeCab를 자체적으로 수정한 한글형태소 분석기를 이용하였다

수집된 텍스트 데이터의 경향을 파악하거나 데이터간의 관계를 가시화 하여 해석가능한 정보로 구체화 하는 방법으로 주로 빈도분석, 연관성분석, 군집분석 등을 사용하고 있다. 연관성 분석은 수학의 그래프 이론에 뿌리를 두고 있으며, 소셜 네트워크 상에서 연결되는 구조 등을 바탕으로 사용자의 영향력을 측정하여, 중심이 되는 사용자나 키워드를 찾는데 주로 활용된다. 군집분석은 비슷한 특성을 가진 개체에서 유사 특성의 군을 발견하는데 사용된다. 본 연구에서는 담화의 핵심이 되는 단어를 추출하기 위한 방법으로 평균적인 문서에서 다수의 빈도로 출현하는 단어와 비교하여 핵심어를 추출하는 TFIDF(Term Frequency-Inverse Document Frequency)분석법을 사용하였다는 점에서 선행연구들과 텍스트 마이닝 방법론의 차이를 가지고 있다.

2. 해수담수화플랜트사업에 관한 선행연구

본 연구에서 대중의 의견을 파악하기 위한 주제로 다루고 있는 해수담수화플랜트사업은 공학적인 기술을 바탕으로 하고 있으며 기술적인 부분에 관한 논점은 실험을 통해 증명이 가능하므로 사회적인 논쟁이 생길 이유가 없는 사안이다. 하지만 해수담수화 시설을 지역에 유치하는 산업적인 주제가 되면서 사회적 이슈가 되고 있다.

수자원의 공급은 공공재의 공급이라는 인식에서 가격을 매겨 판매하여 이익을 내는 산업으로 인식의 전환이 이루어지고 있다고 볼 수 있다. 이러한 인식의 바탕에서 해수담수화 플랜트 시설을 지역에 유치하여 가동하기 위한 사업이 진행되고 있다고 보여 진다. 이에 대한 증거로 해외시장을 겨냥하여 해수담수화 플랜트 기술의 해외진출을 전략적으로 고려하기 위한 목적으로 실시된 양정석, 김일환(2013)의 연구 및 해수담수화 플랜트의 산업화 전망에 대한 연구, 해수담수화 시설의 시장성장률을 바탕으로 역삼투압 방식에서 발생하는 농축수 문제를 해결하기 위한 효율화 측면을 부각하는 해수담수화플랜트 농축수의 자원화에 대한 연구(김승현, 2016), 역삼투압 방식을 기반으로 한 해수담수화 시설의 해외시장전망에 대한 연구(김인수 외, 2009) 등이 이루어지고 있다는 사실을 들 수 있다.

따라서 본 선행연구 검토에서는 해수담수화 시설의 기술적인 문제를 다루는 연구를 제외하고 해수담수화 플랜트 기술을 산업화하여 해외에 진출하기 위한 연구와 지역의 물 부족문제를 해결하기 위한 대안으로 접근하는 연구에 집중하였다.

양정석, 김일환(2013)의 연구는 해수담수화 플랜트 기술의 해외진출을 전략적으로 고려하기 위한 목적으로 해수담수화 플랜트의 시장 진출 유망 국가에 대한 분석을 실시하였다. 수요가 존재하는 11개국을 대상으로 델파이 기법을 이용한 가중치 적용을 통해 미국, 중국, 사우디아라비아 순으로 유망국가지수를 도출하였으나 실제 해외진출은 어려울 것으로 판단하고 있다.

박희경, 박노석(1998)은 해수담수화의 국내 적용을 위한 타당성 검토를 실시하였다. 산업의 성장추세에 따라 공업용수 부족이 예상된다는 전망을 토대로 한다. 효율적인 물 관리를 위해서 댐과 광역상수도 건설 방안이 이미 존재하지만 이를 보완할 대체 수자원이 필요하며, 가뭄대비 측면에서 수자원의 다변화가 요구되는 가운데 해수담수화 단위비용 산정을 통한 경제적 타당성 검토결과 해수담수화 도입방안을 적극 검토해야 된다고 결론짓고 있다.

해수담수화 시설의 표준모델 적용방안에 대한 연구(김충환 외, 2000) 에서도 “인구증가와 공업화의 지속적인 증가”를 근거로 영, 호남지방의 장래의 담수공급부족을 상정하고 있다. 이러한 배경을 토대로 동해의 A지역을 대상으로 역삼투압 방식의 해수담수화 기술을 도입하기 위한 표준모델의 처리량을 추산하는 연구를 진행하였으나, 결론에서는 ”장래의 물 부족에 대비하여 대체 수자원으로 간주할 수는 있으나 기존에 계획된 담수자원의 개발보다 우선하거나 대등할 수는 없을 것” 이라는 표현을 통해 2000년 당시의 기술로는 해수담수화 시설이 담수시설 확충에 비해 효용이 낮다는 점을 시사하고 있다.

해수담수화 플랜트의 산업화 전망에 대한 연구에서는 해수담수화 시설의 시장성장률을 바탕으로 역삼투압 방식에서 발생하는 농축수 문제를 해결하기 위한 효율화 측면을 부각하는 해수담수화플랜트 농축수의 자원화에 대한 연구(김승현, 2016), 역삼투압 방식을 기반으로 한 해수담수화 시설의 해외시장전망에 대한 연구(김인수 외, 2009)가 진행되었다. 이와 같이 산업화와 관련된 선행연구는 역삼투압방식을 적용 개선하는 것을 중심으로 연구가 진행되고 있으며, 기장군 역삼투압 방식 해수담수화 플랜트를 중심으로 한 해수담수화플랜트 시스템 설계 및 시공기술(우성우, 2012)에 관한 연구에서는 기장군의 역삼투압 플랜트를 중심으로 기술향상동향에 대한 연구를 진

행하고 있다.

해수담수화 기술의 미래 시장성에 주목하여, 차세대 해수담수화 기술동향에 대해 설명하고 있는 이상호, 최준석, 황태문(2013)의 연구는 산업화 측면에서 다른 접근을 하고 있다. 현재 시장을 주도중인 기술은 역삼투압 방식이지만, 역삼투압 방식과 증발 방식은 에너지 효율이 낮고, 해수농축수의 후처리 등의 단점들이 많으므로, 향후 10년을 내다볼 때 새로운 고효율의 기술이 대두될 가능성이 높으며, 미래 시장을 선점하기 위해서는 새로운 기술개발에 역량이 투자되어야 한다고 결론짓고 있다.

핵발전소 주변 주민의 건강과 삶의 질에 관한 한상진(2016)의 연구에서는 기장군 해수담수화시설에 관한 사례를 환경정의(Environmental Justice)라는 개념을 통해 분석하고 있는데, 몇 가지 사실을 통해 쟁점을 잘 부각하고 있다. 첫째, 수심 10~15m바닷물을 끌어와 수돗물을 만들 계획이지만 11㎞ 떨어진 지점에 고리 핵발전소가 위치해 있다는 점과 해수담수화물에는 미네랄이 거의 없어 인공적인 화학시약으로 Ca, Mg, Fe 등을 첨가해야 하는 천연미네랄 물과 다름으로(이태호, 2015 재인용) 인하여 잠재적으로 주민들의 신체건강을 위협할 요소가 있다. 둘째, 두산중공업과 중앙정부, 지자체의 합작 프로젝트로 진행되는 과정에서 부산광역시가 건설계획수립 당시에는 식수 공급이 목적이 아니라는 이유로 주민공청회나 설명회를 열지 않았다. 셋째, 해수담수화 수돗물을 톤당 원가는 부산 일반 정수장의 941원보다 비싼 1,187원이며 상수도본부가 5년 후에 시설을 인수 받게 되면 추가로 관리비가 발생될 것이다(이태호, 2015). 넷째, 부산시의 상수도시설의 생산능력은 2015년 기준 하루 209만9천t이지만 이용률은 절반 수준인 53.3%이다. 부산시의 식수공급량은 수요량을 초과하고 있으므로 불필요한 과잉생산이 발생한다. 위의 네 가지 쟁점으로 요약하고 있으며 반대주민들의 이해관계의 초점이 집값하락, 생업피해와 같은 개인의 경제적 사안이 아니라 핵발전소 인근 바닷물에 대한 잠재적 위험에 대한 공포심과 미래세대의 지속가능성을 염려하는 지역사회 차원의 문제로 다루어지고 있다고 문제의 성격을 설명하고 있다.

3. 부산시 해수담수화 시설유치에 대한 쟁점

위와 같은 선행연구의 분석을 통해 부산시 해수담수화 시설유치에 관한 쟁점을 수자원부족의 문제, 해수담수화 시설의 위치상의 적합성 문제, 해수담수화 용수의 생활용수 적합성의 문제, 해수담수화 기술의 경제성 문제, 지역유치정책결정과정의 절차상의 문제 등 다섯 가지 큰 쟁점으로 도출할 수 있다. 이 다섯 가지 쟁점에 대한 세부적인 설명을 추가하여 〈표 1〉에 나타내었다. 긍정적인 측면과 부정적인 측면의 세부사항에 대한 근거는 앞서 선행연구 부분에서 인용과 주석을 통해 모두 제시하였다.

〈표 1〉 선행연구를 통한 부산시 해수담수화 시설유치에 대한 쟁점정리

쟁점	긍정	부정
수자원 부족에 관한 문제	- 한국은 유엔이 선정한 물 부족 국가 (정보의 차이) - 산업용수 및 생활용수가 부족할 전망 (해석의 차이) - 인구증가와 지속적인 공업화 (해석의 차이)	- 한국은 유엔이 선정한 물 부족 국가는 잘못 전달된 정보 (정보의 차이) - 부산 경남권의 상수도 시설현황은 과잉공급 상황 (정보의 차이) - 인구감소와 제조업 해외이탈이 현재 추세 (해석의 차이)
해수담수화 시설 위치상 문제	- 고리원전에서 방출하는 용수에 기본적으로 방사능은 제거됨 (정보의 차이) - 방사능 동위원소는 역삼투압방식을 통해 제거됨 (정보의 차이)	- 원전용수에 방사능이 유출될 가능성이 존재함 (정보의 차이) - 방사성 동위원소인 삼중수소는 역삼투압방식을 통해 제거가 불가능함 (정보의 차이)
생활용수 적합성의 문제	- 생활용수 및 산업용수로 문제없이 공급가능함 (정보의 차이)	- 역삼투압 방식을 거친 용수는 미네랄이 제거됨으로 장기간 식수로 사용할 경우에 인체의 미네랄균형을 무너뜨릴 가능성이 존재함 (정보의 차이)
기술의 경제성 문제	- 용수공급이 부족해질 경우를 대비했을 경우 다양성의 확보가 경제적인 가치를 가짐 (해석의 차이)	- 단위생산가격이 기존의 상수도시설보다 비싸므로 경제적 합리성이 부족함 (정보의 차이)
지역유치결정과정의 절차상의 문제	- 해수담수화 단위비용 산정을 통한 경제적 타당성 검토결과 해수담수화 도입방안을 적극 검토해야 함. (해석의 차이)	- 공공사업의 경우에 계획수립 전에 시민공청회 개최가 정당한 절차로서 지켜져야 하지만 당시에 식수공급목적이 아니라는 이유로 설명회 공청회를 가지지 않았음 (해석의 차이) - 현재는 식수로 공급할 계획으로 변경됨 (해석의 차이)

자료: 필자 작성

선행연구로부터 〈표 1〉과 같은 인식의 차이를 확인하는 것은 서술된 문장을 읽고 맥락을 이해함으로써 가능하다. 해수담수화플랜트 사업에 관한 선행연구의 질적인 분석으로 인식의 차이를 확인할 수 있는 것이다. 그리고 이러한 인식차이의 근거에는 정보의 차이, 해석의 차이(이해관계자의 입장의 차이)가 존재한다. 정보의 차이에 따른 인식의 불일치는 정확한 데이터가 있다면 쉽게 해소 가능한 속성을 가지지만 모두가 납득 가능한 정확한 정보를 획득하기 어려운 문제가 존재한다. 해석의 차이에 따른 인식의 불일치는 각자 속한 입장에서 의도적, 전략적인 발언의 속성을 가지게 됨으로 합의점을 찾아가는 과정이 어렵다. 따라서 공적토의를 분석하는데 있어서 발생하는 인식차이가 정보의 차이에 기인하는 것인지 해석의 차이에 기인하는 것인지를 파악하는 것은 대단히 중요하다.

수자원 부족에 관한 인식의 차이는 “한국은 유엔이 정한 물 부족 국가”라는 잘못된 정보가 전파되고 있음으로 해서 나타난다. 따라서 “한국이 유엔이 정한 물 부족 국가”가 잘못된 정보라는 사실을 인지하게 되면 대립 중에 있던 의견의 차이는 해소된다. 하지만 인구증가와 지속적인 공업화가 미래에 발생할 것이라는 입장에서 수자원 확충을 주장하는 쪽과 인구감소와 제조업이탈이 미래에 발생할 것이라는 입장에서 수자원 과잉을 주장하는 쪽의 의견은 미래에 대한 자신의 해석을 토대로 하고 있기 때문에 의견의 차이를 좁히기 어려운 쟁점이 된다. 그러나 상대방의 입장에서 상반된 의견에 대한 근거를 서로 검토함으로써 합의점에 도달할 수 있는 가능성이 생기고 이러한 인식

을 차이를 좁히기 위해 미시적인 회의석상에서는 대화와 토론을 실시하게 된다.

거시적 토의는 미시적 토의와 달리 단순 대화를 통해 맥락을 이해하는 것이 불가능하다. 거시적 토의에서는 서로 상반된 입장들이 정보의 차이에서 발생하는 것인지, 혹은 입장의 차이에서 나타나는 것인지를 분류하기 위한 방법론이 요구되며, 이러한 방법론을 통한 분석이 필요하다. 더불어 하나의 공통된 핵심주제를 가지고 토론이 제대로 이루어지고 있는지에 대한 분석이 필요로 된다. 텍스트를 정독하거나 상대의 의견을 경청하는 방법으로는 담론의 일부만을 세부적으로 다루게 되는 등의 한계점을 갖는다. 따라서 본 연구는 기존 분석의 한계점을 극복하고 토의 전체 내용을 자동화하여 분석할 수 있는 코퍼스 분석기법을 사용하여 연구를 진행하였다.

Ⅲ. 형태소분석 및 TFIDF분석

1. 데이터 수집

“부산 해수담수화 플랜트사업”에 관한 거시적인 토의에서 각 미디어 특성에 따른 문제인식과 의견차이가 있는지를 파악하기 위하여 지방일간지 2개사와 중앙일간지 3개사를 선정하였다. 이후 “부산, 해수담수화” 키워드로 검색한 결과물을 월별로 분류하여, 기사제목, 본문, 작성자 등을 수집하였다(〈표 2〉참조). 각 일간지 웹사이트의 검색결과에 나타나지 않은 수집상의 누락은 일부 존재할 수 있으며, 부산지역과 관련 없는 해수담수화에 관한 기사도 기술적인 문제의 쟁점을 다루고 있는 경우가 대부분이므로 수집대상에 포함하였다.

〈표 2〉 신문 미디어 5개사 자료수집

분류	성격	시기	데이터 수
미디어 1	지역일간지	2015년 9월 – 2016년 6월	92
미디어2	지역일간지	2015년 7월 – 2016년 6월	92
미디어3	중앙일간지	1995년 2월 – 2015년 12월	23
미디어4	중앙일간지	2014년 12월 – 2016년 5월	14
미디어5	중앙일간지	2009년 1월 – 2016년 3월	12

지역일간지는 평균 92건의 기사를 보도하였고, 중앙일간지는 평균 16건의 보도가 있었다. 미디어 3 사의 경우 조사대상인 언론사 중 가장 빠른 10년 전인 1995년에 부산지역에 해수담수화 플랜트 시설이 요구될 가능성이 있다는 요지의 기사를 보도했으며, 다른 언론과는 달리 논란이 한창인 2016년에는 보도기사가 존재하지 않고, 2015년의 기사가 부산지역 해수담수화에 관한 마지막 보도였다.

2. 형태소 분석

대량의 비정형텍스트 데이터를 전제로 하는 텍스트 마이닝 기법은 자연어 처리기술을 토대로 이루어진다. 비정형데이터는 고정된 필드에 저장되어 있지 않고 분산된 장소에 각기 다양한 형태로 존재하는 정형화 되지 않은 데이터이다. “부산, 해수담수화” 키워드로 검색한 233개의 지방일간지 및 중앙일간지의 보도기사로 이루어진 비정형텍스트 데이터의 문장으로부터 자연어처리를 위한 형태소분석을 실시하였다. 형태소분석 과정은 현대한국어의 형태소 분석이 가능하도록 수정한 오픈소스 형태소분석기 MeCab를 사용한 어노테이션(Annotation)을 통해 이루어졌다. 어노테이션은 텍스트 덩어리를 단어의 “명사”, “조사”, “형용사”, “동사” 와 같은 형태소별로 분해하여 각 형태소 단위로 주석을 다는 것을 의미한다.

이를 위해 언어, 사전 코퍼스에 의존하지 않는 범용적인 설계를 기본방침으로 제작된 MeCab에 현대한국어 텍스트를 형태소로 분석하여 형태소정보를 부여하기 위해 만들어진 분석사전인 HanDic을 이용하여 현대한글형태소 분석이 가능하도록 수정하였다. 본 연구에 사용된 형태소분석기의 결과 값은 명사의 경우 고유명사(이름, 도시 등), 의존명사, 보통명사, 조사의 경우 주격, 속격, 구격 등을 구분하고, 접두사, 접미사를 구분한 주석 달기를 통해 기계적인 분석이 가능한 메타텍스트 데이터로 변환되어 나타난다.

최초 수집한 233개의 일간 보도자료를 텍스트 파일로 저장하였다. 그 뒤 하루 단위로 저장되어진 텍스트 자료를 월단위의 시계열로 분석하기 위하여 각 미디어 별, 월 별로 합쳐서 하나의 파일로 저장 하였다. 그 결과 미디어1의 경우 9개, 미디어2는 10개, 미디어3은 6개, 미디어4는 5개, 미디어5는 9개의 재구성한 텍스트 파일을 생성하였다. 즉, 총 233개의 일간 보도자료를 39개의 미디어별 월별 텍스트 덩어리로 재구성하였다. 결국 분석에 사용된 텍스트의 총량은 233개의 일간 보도자료의 문자수와 같다. 분석은 텍스트를 XML을 기초로 한 GDA(Global Document Annotation) 파일 형태로 저장하여 두 번에 걸쳐서 진행하였다. 233개의 하루단위의 텍스트를 미디어별로 묶어서 5개의 텍스트 덩어리로 분석을 실시한 뒤, 시계열 분석을 위해여 39개의 월별 텍스트 덩어리로 나누어 분석을 실시하였다.

〈그림 1〉 데이터 수집 및 분석순서

3. TFIDF분석

대용량 코퍼스의 자동화된 분석과정에서 핵심단어추출의 정확성을 높이기 위해 사용되는

TFIDF법(Salton and McGill, 1983)은 문서에서 단어수의 중요도를 판정하여 중요하다고 판정되는 언어만 분석의 요소로 사용하는 방법이다. 본 연구에서는 수집된 미디어 자료를 형태소 분석을 통해 전 처리한 GDA메타텍스트를 기초로 하여 각 미디어별로 중심인식(Major Opinion)에 가까운 단어를 추출하기 위해 TFIDF분석을 실시하였다.

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

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

$TF_{w,a}$ = Number of occurences of term w in document a

DF_w = Number of documents containing term w

N= Total number of documents

$TFIDF_{w,a}$ 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 단어 빈도수를 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_w의 로그지수로 얼마나 많은 문서에 단어 w가 출현하는지를 나타내는 것이다. 예를 들어, 만약 단어 w1이 비교문서에서 적은 수로 나타나고, 분석문서에서는 많은 수로 나타난다면, w1의 IDF 값은 높고, TFIDF스코어도 높아진다. 이와는 반대로, 어떤 단어 w2가 비교문서에서 많이 나타나지만, 분석문서에는 적게 나타난다면, 단어 w2의 IDF값은 낮아지고, TFIDF스코어 또한 낮아진다. 〈식 1〉에서 나타내는 것처럼 IDF값과 TF값이 높으면 높을수록, TFIDF값이 높다. j문서의 i단어의 수를 예로 들면 TFIDF값이 높은 i단어는 j문서에서 중요한 단어이며, 담화의 핵심어라고 할 수 있다.

본 연구의 IDF(Inverse Document Frequency) 비교문헌 값으로는 카이스트가 제공하는 620건의 신문데이터를 이용하였다. 카이스트가 제공하는 IDF 샘플용 신문 데이터는 미디어의 토픽추출을 위해 작성 된 샘플로 상투적인 언어를 토대로 하고 있다.

부산시 해수담수화플랜트사업에 관한 미디어 5개사의 메타데이터 TFIDF 분석결과는 〈표 3〉과 같다. 〈표 3〉에 제시한 결과를 살펴보면, 데이터 샘플 수가 많을수록 TFIDF값이 높게 나타나는 핵심어의 추출 빈도도 높게 나타났다. 일반적으로 생각해 볼 때 문서의 수가 많아지면, 단어의 수도 많아지는 것은 당연하다. 하지만 하나의 논점을 논의하는데 있어 사용되는 단어수가 많다는 것은, 표현의 다양성과 더불어, 논점의 다양성이 존재할 가능성이 크다는 것을 의미한다. 즉, 서로 다른 관점과 인식이 표현되고 있을 가능성이 있다는 것을 의미한다. 예를 들어 미디어 1과 2는 미디어 3,4,5와 비교하여 더욱 다양한 관점과 인식을 가지고 있다고 해석할 수 있다.

〈표 3〉 TFIDF 분석결과 개요

구분	미디어 1	미디어2	미디어3	미디어4	미디어5
샘플 수	92	92	23	14	12
TFIDF추출단어	3204	3410	1359	813	1230

각 미디어별로 중요한 인식 지표를 나타낸다고 할 수 있는 TFIDF 추출결과로부터 각 미디어가 중요하게 사용하는 단어가 무엇인지 알 수 있다. “해수”라는 단어는 주제자체인 단어이기 때문에 5사가 공통으로 가장 높은 TFDF값을 나타내고 있다. 미디어 1, 2, 4사에 두 번째로 나오는 “주민”이라는 단어는 해수담수화 플랜트사업과 관련된 가장 중요한 핵심단어이다. 미디어 1, 2, 4사는 해수담수화 플랜트사업에 있어서 “주민”이라는 단어를 가장 많이 사용하고 있다. 그러나 미디어 5의 경우에 “해양”, “공급”이라는 단어 다음에 “주민”은 네 번째로 나타났다. 미디어 3의 경우에는 “상수도, 공급, 시설, 해양, 바닷물, 정수, 센터, 생산, 기술, 사업, 설비” 등의 용어 뒤에 열세 번째 단어로 “주민”이 나타나고 있다. 이러한 결과 값은 미디어 3사의 핵심단어 사용이 다른 미디어 4개사와 비교했을 때 상대적으로 다르게 나타나고 있다는 사실을 나타낸다. 미디어 1, 2, 4, 5사는 “원전”, “방사”, “방사능”이라는 단어가 핵심단어로 다루어지고 있다. 하지만 미디어 3에서는 “방사”와 “방사능”은 상위 50개 단어 안에서 다루어지고 있지 않으며, “원전”이 43번째 순으로 나타나 매우 낮은 TFIDF값을 보였다.

이와 같이 핵심단어들의 분석결과들을 종합하면 언어사용의 경향이 드러나게 된다. 핵심단어를 통해 종합된 언어사용의 경향을 토대로 미디어 5개사의 인식의 차이라고 할 수 있는 다음과 같은 시사점을 도출해낼 수 있다. 지역일간지와 중앙일간지간의 차이는 기사건수의 차이에서부터 현안으로서 중요하게 다루어지는 온도차이가 존재한다. 기사 내용의 분석을 통해 알 수 있는 문제인식의 차이 또한 뚜렷하다. 중앙일간지 3개사간의 인식의 차이는 가장 뚜렷한 점이 논란의 인정여부에서 드러나고 있다. 미디어4와 미디어 5에서는 “반대”, “대책위”, “토론회”, “철회”, “검사” 등의 단어 사용을 통해 반대의견이 존재하고 논란이 되는 이슈임을 인식하고 있지만, 미디어 3은 이러한 의미를 내포하는 뉘앙스를 가진 단어를 한 단어도 사용하지 않음으로써 부산의 해수담수화 사업에서 갈등이나 반대의견이 마치 존재하고 있지 않는 것과 같은 입장을 취하고 있다는 사실을 알 수 있다.

Ⅳ. 시계열 TFIDF분석

1. 시계열분석을 위한 단어의 의미단위화

앞선 데이터 전체의 TFIDF분석을 통해 알 수 있는 시사점을 토대로 지방 2개사와 중앙일간지 3대 신문사의 토픽의 변화를 파악하기 위하여 시계열 TFIDF분석을 실시하였다. 각 신문사가 해수

담수화를 기사로 다루면서 중점적으로 사용하는 핵심어의 변화를 통해 내용전달의 방향성이 시간이 흐름에 따라 어떻게 변화해 가는지를 살펴보기 위한 분석이다. 5개 신문사별 핵심어의 시계열 변화의 차이를 살펴봄으로써 관점의 차이를 더욱 명확히 파악할 수 있으며, 이는 여론형성과정에 각 미디어가 어떻게 영향을 미치는지를 분석하기 위한 기초자료로 활용할 수 있다.

핵심어의 변화를 파악하는 지표로 사용하기 위하여 시계열로 추출된 단어들을 사용하는데 있어서, 수백단위의 단어를 하나하나 개별적으로 분석 하는 것은 시계열 분석에서 변화흐름을 보기위한 분석으로는 용이하지 않으며, 결과에 대한 해석을 어렵게 한다. 따라서 단어의 의미단위(Conceptual Unit)로 추출단위의 군집화를 실시하여 시계열 분석을 실시하였다. 의미단위의 분류는 선행연구 분석과 전체자료의 TFIDF분석을 통해서 파악된 시사점들을 토대로 해수담수화 공장의 부산지역유치 및 상수도 공급이라는 정해진 맥락을 반영하여 나타내는 의미를 군집화하여 분류하였다.

〈표 4〉와 같이 크게는 위험요소와 갈등요소로 분류되어진 단어군에서, 위험요소를 3개 군으로 나누어 원자력에 대한 위험을 나타내는 의미단위, 질병에 대한 위험을 나타내는 의미단위, 기타 위험요소에 대한 의미단위로 세분화 하였다. 또한 세분화된 의미단위를 의미단위 1~4로 각각 분류하여 시계열 분석을 실시하기 위해서 각 단어의 시기별 다른 기사 수와 단어 수를 감안한 표준치를 정의하였다.

〈표 4〉 단어의 의미단위(Conceptual Unit) 분류

구분	개념	단어
의미단위 1	위험요소-원자력	원자력, 핵연료, 원전, 방사(능), 폐기물
의미단위 2	위험요소-질병	암, 발암, 유방암, 증후군, 질병, 스트레스, 발병률, 후유증, 발생률, 사산, 피로감, 악화, 유해, 치명적, 해악
의미단위 3	위험요소-기타	환경, 재난, 생명, 생태, 오염, 지진, 사고, 고장, 파괴, 노후, 위험, 밀집, 열악
의미단위 4	갈등요소	갈등, 규탄, 방해, 비난, 반발, 반대, 부적절, 핑계, 이기주의, 분노, 무책임, 과시, 무시, 불통, 허술, 비판, 불신, 거짓, 불만, 부정, 엉터리, 날림, 복잡, 희생, 슬픔, 위법, 과도, 분통, 맞불

2. 시계열분석을 위한 단어의 표준치 정의

각 신문에서 추출된 단어수가 시기별로 각각 다르기 때문에, TFIDF값을 그대로 적용하여 시계열 분석을 할 수 없다. 따라서 본 분석을 실시하는 과정에서 TFIDF 값으로 정렬된 단어의 순위를 기준으로 표준치z를 아래 식과 같이 정의하여 시계열 분석을 실시하였다.

$$z = 1 - \frac{r}{n} \quad \langle \text{식 } 2 \rangle$$

z = 표준치, r = TFIDF순위, n = 해당 신문에서 추출된 단어의 수

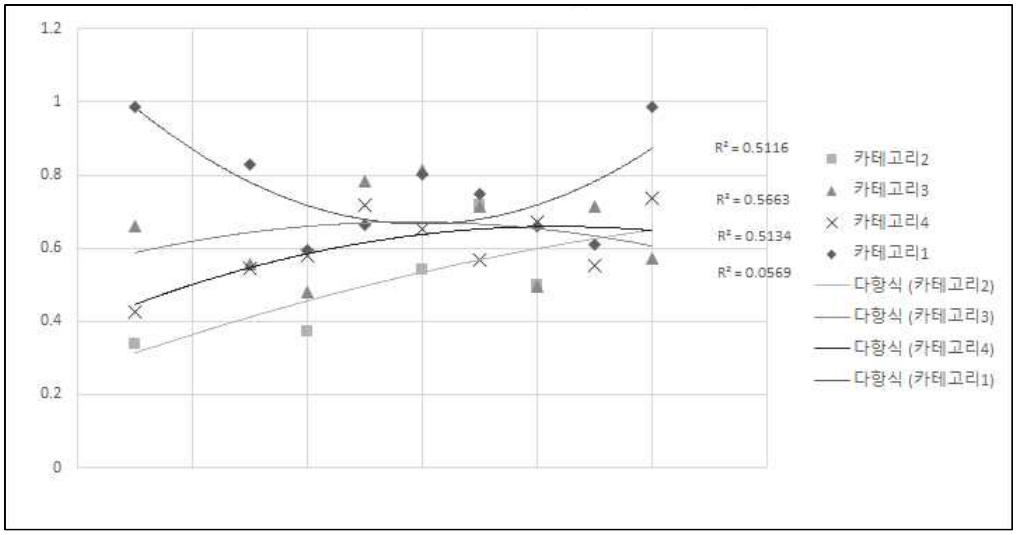
〈표 5〉 단어별 z표준치 (미디어 1의 코퍼스 예, 소수점 이하 3자리까지)

의미단위 1	2015년 9월	2015년 10월	2015년 11월	2015년 12월	2016년 1월	2016년 2월	2016년 3월	2016년 4월	2016년5 5월	2016년6 6월
원자	-	-	-	0.069	-	-	-	-	-	-
원자력	-	-	-	-	0.577	0.971	-	0.443	0.528	-
핵연료	-	-	-	-	-	-	0.607	0.636	-	-
원전	0.985	-	0.940	0.979	0.641	0.997	0.988	0.983	0.881	0.985
방사	-	-	0.630	0.971	0.923	0.977	0.567	0.576	0.959	-
방사능	-	-	0.912	0.926	0.875	0.991	0.836	-	-	-
폐기물	-	-	-	0.037	0.304	0.070	-	-	0.073	-
z	0.985	-	0.828	0.596	0.664	0.801	0.75	0.660	0.610	0.985

3. 신문사별 코퍼스 변화 (z표준치의 시계열분석)

신문사별 코퍼스 변화를 z표준치의 시계열분석을 통해서 살펴볼 수 있다. 다만, 여론이나 사회적 현상에 민감하게 반응하는 신문이라는 미디어 데이터의 특성상, 신문사에 따라서 또는 특정시기에 따라서 당해분석과 관련된 코퍼스 데이터가 존재하지 않는 범위가 존재한다. 이러한 범위의 값을 예측하는 통계분석방법의 하나로 회귀분석이 있다. 즉, 데이터의 추세를 계측하여 데이터가 존재하지 않는 범위를 예측하는 방법이다. 신문코퍼스는 데이터의 변동이 심하므로, 본 연구에서는 다항식 추세선을 이용하여 시간에 따른 코퍼스의 변화를 예측하였다.

〈그림 2〉 미디어 1의 코퍼스 변화 (다항추세선)

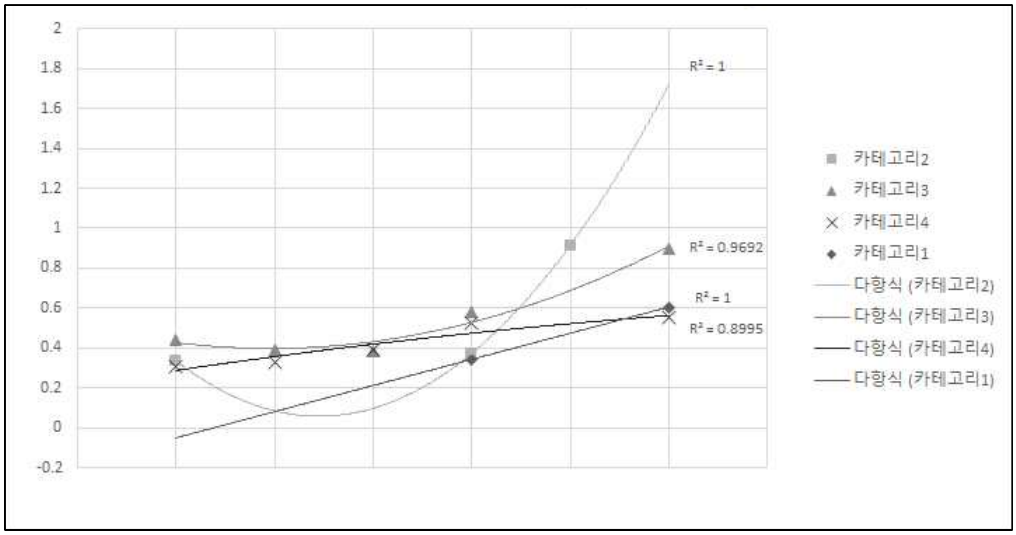


〈그림 2〉는 미디어 1의 코퍼스 변화를 예측하기 위하여, z표준치의 다항추세선을 표기한 것이

다. 의미단위 1 (위험요소-원자력)의 경우는 오목형의 변화를 보인다. 즉, 시기를 초기, 중기, 말기로 나누어 보면, 초기에는 의미단위1에 대해 많이 언급하다가, 중기로 갈수록 언급을 자제하고, 말기에서 다시 자주 언급하는 것을 알 수 있다. 또한, z표준치의 범위는 0.5와 1사이로 다른 의미단위에 비하여 비교적 높은 값을 보인다. 여기서, 결정계수 R2 값은 추세선의 예상 값이 실제 데이터와 일치하는 정도를 나타낸다. 0~1 사이의 값을 가지며, 1 에 가까울수록 가장 안정적이며, 0.5 이상이면 정밀도가 높은 편이라 할 수 있다. 즉, 미디어 1 코퍼스의 논점이 다른 이슈와 비교하여, "원자력"에 중점을 두고 있으며, 언급의 정도에 변화는 있지만, 앞으로 더 많이 언급하는 방향으로 변화하고 있음을 알 수 있다. 또한, 의미단위 3 (위험요소-기타), 의미단위 2 (위험요소-질병), 의미단위 4 (갈등요소)는 약한 볼록형의 변화를 보인다. 언급정도에 큰 변화는 없으나 일정정도 중요하게 다루고 있는 것으로 나타난다. 또한, z표준치의 범위를 기준으로 살펴보면, 0.3에서 0.6사이의 범위 내에서 변화하고 있으며, 의미단위 3, 4, 2의 순으로 중요하게 다루고 있는 것을 알 수 있다.

미디어 1의 코퍼스 변화를 기준으로 하여 나머지 4개 신문사의 분석결과를 비교하였다. 신문사별 코퍼스의 시계열에 따른 변화차이를 살펴봄으로써 관점의 차이를 더욱 명확하게 파악할 수 있다. 이는 여론형성 과정에 각 미디어가 어떻게 영향을 미치는 지를 분석하기 위한 기초자료로 활용할 수 있다.

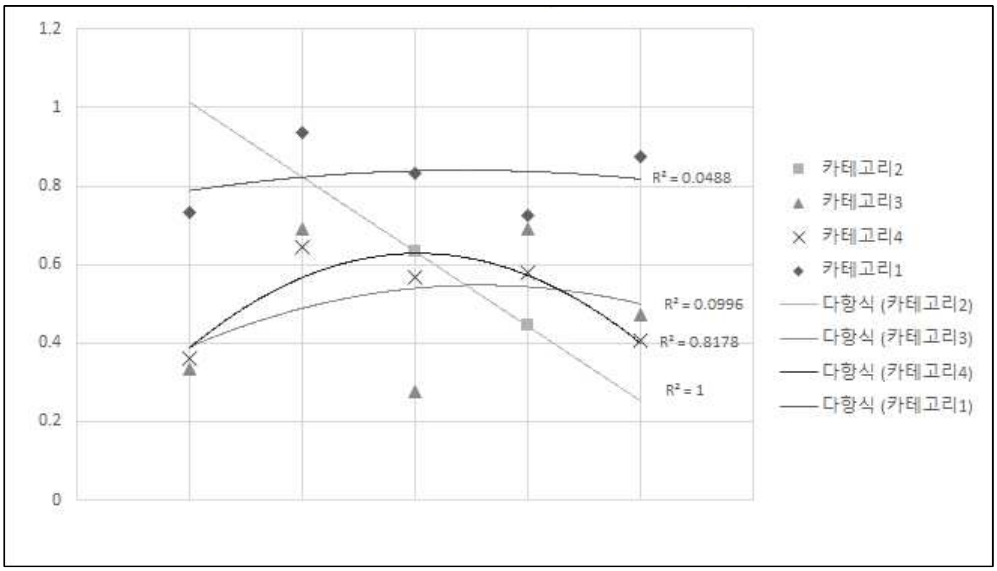
〈그림 3〉 미디어 3의 코퍼스 변화 (다항추세선)



〈그림 3〉은 미디어 3의 코퍼스 변화를 살펴보기 위하여, z표준치의 다항추세선을 표기한 것이다. 의미단위 1 (위험요소-원자력), 의미단위 3 (위험요소-기타), 의미단위 4 (갈등요소)의 경우는 직선상향의 변화를 보인다. 이는, 시간흐름에 따라 언급의 정도가 많아지고 있음을 나타낸다. 또한, z표준치의 범위 살펴보면, 의미단위 2 (위험요소-질병), 의미단위 3 (위험요소-기타), 의미단위 4 (갈등요소), 의미단위 1 (위험요소-원자력)순으로 중요하게 다루고 있는 것을 알 수 있다. 특히,

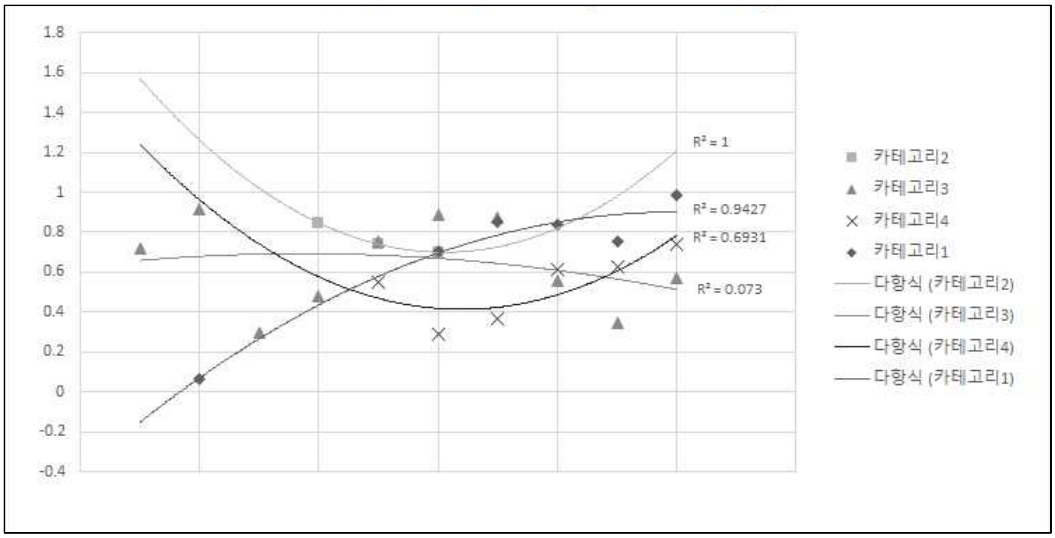
의미단위 1 (위험요소-원자력)에 대해서는 z표준치의 범위가 직선상향변화를 보이기는 하지만, 0과 0.5사이로 다른 의미단위에 비하여 비교적 낮은 값을 보인다. 이는 미디어 3 코퍼스의 논점이 다른 이슈와 비교했을 때, "원자력"에 중점을 두지 않고 있음을 보여준다. 의미단위 2 (위험요소-질병)의 경우, 오목형의 상향변화를 보인다. 즉, 시간흐름에 따라 언급의 정도에 큰 차이가 있으며, z표준치의 범위가 0에서 1.5사이의 범위로, 미디어 3 코퍼스의 논점이 다른 이슈와 비교하여, "질병"에 중점을 두고 있음을 알 수 있다.

〈그림 4〉 미디어 4의 코퍼스 변화 (다항추세선)



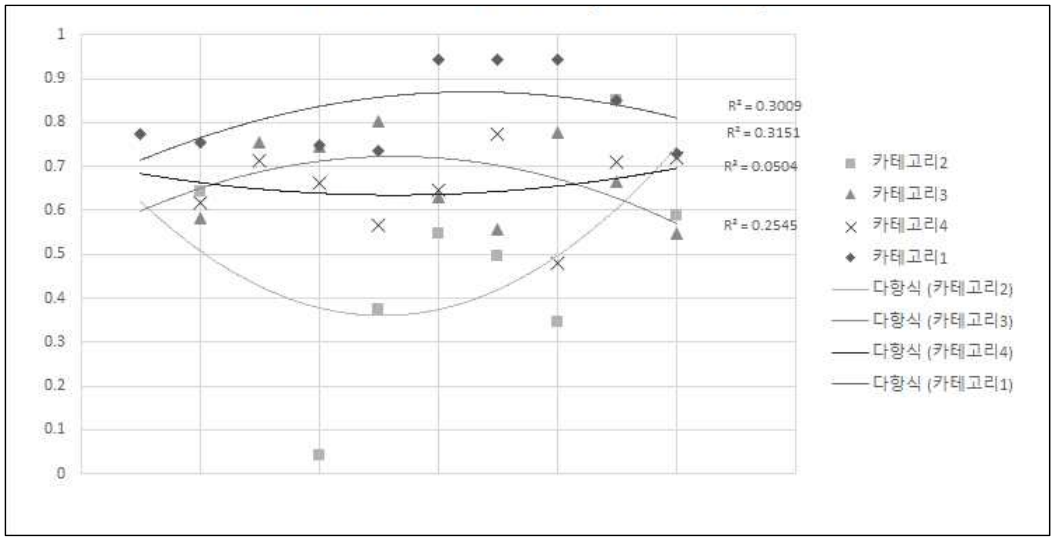
〈그림 4〉는 미디어 4의 코퍼스 변화를 면밀히 살펴보기 위하여, z표준치의 다항추세선을 표기한 것이다. 의미단위 1 (위험요소-원자력)의 경우는 직선에 가까운 약한 볼록형의 변화를 보이고 있다. 이는 시간흐름에 따라 언급의 정도에 차이가 거의 없음을 의미한다. 또한, z표준치의 범위는 0.8정도로 다른 의미단위에 비하여 상당히 높은 값을 보인다. 즉, 미디어 4 코퍼스의 논점이 다른 이슈와 비교하여, "원자력"에 지속적으로 중점을 두고 논의하고 있음을 알 수 있다. 의미단위 3 (위험요소-기타), 의미단위 4 (갈등요소)는 비교적 강한 볼록형의 변화를 보여 언급정도에 차이가 있음을 보여주고 있다. 또한, z표준치의 범위는 0.4에서 0.6으로 의미단위 1 (위험요소-원자력)에 비해서는 낮은 범위이긴 하지만, 꾸준히 중요하게 다루고 있음을 알 수 있다. 의미단위 2 (위험요소-질병)에 대해서는 하향직선형으로, z표준치의 범위는 1에서 0.3으로 언급의 정도가 점점 떨어지고 있음을 알 수 있다.

〈그림 5〉 미디어 5의 코퍼스 변화 (다항추세선)



〈그림 5〉는 미디어 5의 코퍼스 변화를 면밀히 살펴보기 위하여, z표준치의 다항추세선을 표기한 것이다. 의미단위 2(위험요소-질병)와, 의미단위 4(갈등요소)는 오목형의 변화를 보인다. 이는 시간흐름에 따라 언급의 정도에 차이가 있음을 나타낸다. 또한, z표준치의 범위는0.3에서 1사이로 시간에 따라 큰 변화를 보인다. 언급을 많이 하는 시점에 있어서 어떤 특정한 경제적, 정치적, 사회적 사건에 영향을 받은 것으로 추정된다. 언급을 많이 하는 시점이나 가장 적게하는 시점의 전후에 대한 코퍼스의 내용분석을 필요로 하는 부분이다. 의미단위 1 (위험요소-원자력)의 경우는 상향직선에 가까운 볼록형의 변화를 보여 시간흐름에 따라 언급의 정도가 높아지고 있음을 나타낸다. 또한, z표준치의 범위는 0과 1사이로 다른 의미단위에 비하여, 시간에 따라 큰 변화를 보인다. 미디어 5 코퍼스의 논점이 다른 이슈와 비교하여, "원자력"에 중점을 두는 방향으로 변화하고 있음을 알 수 있다. 의미단위 3 (위험요소-기타)은 직선에 가까운 볼록형의 변화를 보이며, z표준치의 범위는 0.5로 나타나 일정한 정도로 중요하게 다루고 있는 것을 알 수 있다.

〈그림 6〉 미디어 2의 코퍼스 변화 (다항추세선)



〈그림 6〉은 미디어 2의 코퍼스 변화를 면밀히 살펴보기 위하여, z표준치의 다항추세선을 표기한 것이다. 의미단위 1 (위험요소-원자력)과, 의미단위 3 (위험요소-기타)의 경우는 약한 볼록형의 변화를 보여 시간흐름에 따라 언급정도에 큰 차이가 없음을 나타낸다. 또한, z표준치의 범위는 0.6과 0.8사이로 다른 의미단위에 비하여 상당히 높은 값을 보인다. 즉, 미디어 2 코퍼스의 논점이 다른 이슈와 비교하여, "원자력"과 "기타 위험요소"에 지속적으로 중점을 두고 논의하고 있음을 알 수 있다. 의미단위 4 (갈등요소)는 약한 오목형의 변화를 보여 시간흐름에 따라 주목도에 큰 차이가 없음을 나타낸다. 또한, z표준치의 범위는 0.6과 0.8사이로 다른 의미단위에 비하여 상당히 높은 값을 보여 미디어 2 코퍼스의 논점이 다른 이슈와 비교하여, "갈등요소"에 대해서도 지속적으로 중점을 두고 논의하고 있음을 알 수 있다. 의미단위 2 (위험요소-질병)는 강한 오목형의 변화를 보이고 있는데, 이는 시간흐름에 따라 주목도에 큰 차이가 있음을 의미하는 것이다. 또한, z표준치의 범위는 0.4과 0.7사이로 다른 의미단위에 비해서는 낮은 범위이긴 하지만, 일정 정도 중요하게 다루고 있음을 알 수 있다.

〈표 6〉 신문사별 z표준치 다항추세선의 특성

	미디어 1	미디어 3	미디어 4	미디어 5	미디어 2
의미단위 1 (위험요소-원자력)	오목형* z: 0.59-0.98	직선상향형 z: 0-0.6	약한볼록형* z: 0.72-0.93	직선상향형 z: 0.06-0.98	약한볼록* z: 0.72-0.94
	다른 이슈와 비교하여, "원자력"에 중점을 두고 있으며, 더 많이 언급하는 방향으로 변화	전반적으로 거의 언급이 언급이 없으며 후기에 가서 낮은 z 표준치를 보이며 등장함	직선에 가까운 변화로 시간의 흐름에 따른 변화없이 지속적으로 중요하게 다루고 있음	상향직선에 가까운 볼록형, 시간흐름에 따라 언급의 정도가 높아지고 있음	시간의 흐름에 따른 변화 없이 지속적으로 중요하게 다루고 있음
의미단위 2 (위험요소-질병)	약한볼록형 z: 0.33-0.71	오목상향* z: 0.33-0.91	하향직선형 z: 0.63-0.43	오목형* z: 0.70-0.84	약한오목형 z: 0.04-0.84
	가장 약한 중요도를 내타내지만 언급정도에 큰 변화는 없이 일정정도를 유지하며 다루고 있음	시간의 흐름에 따라 언급의 정도가 큰 차이로 늘어 남, "질병" 요소를 후기에 와서 중점을 두고 언급	언급의 정도가 점점 떨어지고 있음	시간흐름에 따라 언급의 정도에 줄어들었다가 다시 늘어나고 있음	지속적으로 중요하게 다루고 있음
의미단위 3 (위험요소-기타)	약한볼록형 z: 0.47-0.81	직선상향형 z: 0.38-0.89	강한볼록형 z: 0.27-0.69	약한볼록형 z: 0.29-0.8	약한볼록형 z: 0.54-0.80
	두 번째로 중요하게 다루고 있음, 언급정도에 큰 변화는 없으나 일정정도 중요하게 다루고 있음	시간의 흐름에 따라 언급의 정도가 큰 차이로 늘어 남	낮은 범위지만 꾸준히 중요하게 다루고 있음, 언급의 정도는 늘어났다가 점점 떨어지고 있음	추세선은 직선에 가까운 볼록형으로 일정한 정도로 중요하게 다루고 있음	지속적으로 중요하게 다루었으나 후기에 언급이 줄어들고 있음
의미단위 4 (갈등요소)	약한볼록형 z: 0.42-0.73	오목형 z: 0.30-0.55	강한볼록형 z: 0.36-0.64	오목형 z: 0.29-0.73	강한오목형 z: 0.48-0.71
	세 번째 중요도를 나타내나 언급정도의 변화없이 일정한 언급정도를 유지하며 다루고 있음	시간의 흐름에 따라 언급의 정도가 약하게 늘어 남	낮은 범위지만 꾸준히 중요하게 다루고 있음, 언급의 정도는 늘어났다가 점점 떨어지고 있음	시간흐름에 따라 표준치(언급의 정도)가 0.29까지 줄어들었다가 늘어나고 있음	어느 시점에서는 중요하게 다루어지고 어느 시점에는 다루어지지 않는 큰 주목도의 변화를 나타냄
특징	- 원자력를 중점으로 모든 요소들이 지속적으로 언급되며 고르게 나타나는 경향을 보임	- 표준치의 범위 살펴보면, "질병", "기타", "갈등요소", "원자력"순 - "질병"에 중점을 두고 있음 - 초기부터 시작하여 긴 시간 동안 원자력의 언급이 없었음	-"원자력에"에 대해 지속적으로 중심을 두고 언급하고 있으며 "갈등요소"와 "기타위험요소"를 적은 수이지만 꾸준히 언급하고 다루고 있음	- "원자력"의 표준치의 범위는 0과 1사이, 시간에 따라 큰 변화를 보이며 "원자력"에 중점을 두는 방향으로 논점이 변화하고 있음	-"원자력"과 "갈등요소", 생태, 환경 등을 의미하는 "기타위험요소"를 골고루 중심논점으로 다루고 있음

〈표 6〉은 신문사별 z표준치의 다항선추세선의 특성을 정리한 것이다. 여기서 *마크는 각 신문

사가 언급하는 정도가 상대적으로 가장 높은 의미단위를 표시한 것으로, 각 신문사가 중요하게 생각하는 키워드라고 생각할 수 있다. 미디어 1, 미디어 4, 미디어 2은 의미단위 1(위험요소-원자력)에 대한 언급의 정도가 높으며, 해수담수화 논의에 있어서 중요한 키워드로 다루고 있음을 알 수 있다. 반면, 미디어 3과 미디어 5는 의미단위 2(위험요소-질병)에 대한 언급의 정도가 높으며, 해수담수화 논의에 있어서 중요한 키워드로 다루고 있음을 알 수 있다. 결론적으로 해수담수화 논의에 있어서 미디어간 인식의 차이가 존재하는 것을 알 수 있다. 특히, 지역일간지와 중앙일간지의 관점에 명백한 차이가 있음을 분명히 알 수 있다.

미디어를 통한 거시적 담화에 있어서 인식의 차이가 존재 할 경우, 사회적 합의를 이끌어 내는 것은 용이하지 않다. 해수담수화 논의에 대한 인식의 차이가 앞으로 더욱더 극명화 될 수 있으며, 사회적 합의를 이끌어 내기 위해서는 다양한 이해관계자들 간의 심도 있는 논의가 더욱 많이 이루어져야 한다. 또한 미디어의 보도가 이해관계자들의 논의 내용을 객관적으로 투명하게 전달하고 있는지에 대한 평가가 필요하다.

V. 결론: 정책적 함의 및 과제

다양한 이해관계자들이 관여하는 공공사업에서 완전한 의견 일치는 존재하지 않을지도 모른다. 하지만 공공사업을 통해 최대한의 사회적 이익을 달성하기 위해서는 인식을 공유하고 대화와 의견조절을 통해 사회적인 합의점을 찾아가는 과정이 필요하다. 거시적 토의에서 분명히 존재하는 인식차이를 발견하고 좁혀가기 위해서는 우선 인식차이의 정도를 객관화하는 과정이 필요하다.

이에 대한 시도로서 본 연구는 “부산 해수담수화플랜트사업”을 대상으로 미디어간에 존재하는 인식차이를 객관화하기 위하여 시계열로 핵심어를 추출하는 자동화된 오피니언 마이닝 기법을 구축하는 것을 통해 대립이나 갈등을 해소할 수 있는 단서를 제공하는 것을 목적으로 하였다. 지역 2개 신문사와 중앙 3개 신문사의 기사를 수집하여 각 신문사의 해수담수화 플랜트 건설에 대한 비정형 텍스트 데이터로부터 기계적인 분석방법을 통해 인식과 의견을 단순화 하여 나타내는 핵심어를 추출하는 TFIDF를 통한 코퍼스(corpus) 분석을 실시하였다.

“부산 해수담수화플랜트사업”에 대한 분석은 각 미디어사간의 인식의 차이라고 할 만한 뚜렷한 경향성을 보여주었다. 지역언론과 중앙언론간에 중요하게 다루는 정도의 차이가 존재했으며, 미디어 1, 미디어 4, 미디어 2는 문제의 원인이라고 할 수 있는 의미단위 1(위험요소-원자력)을 해수담수화 논의에 있어서 중요한 키워드로 다루고 있는 반면 미디어 3과 미디어 5는 문제의 결과라고 할 수 있는 의미단위 2(위험요소-질병)에 대한 언급의 정도가 높나 해수담수화 논의에 있어서 중요하게 다루고 있음을 알 수 있다.

관심과 중요도가 커져가고 있는 코퍼스 분석 분야에 있어서 본 연구의 방법론으로 사용된 한글 자연어처리 기술과 TFIDF분석을 통한 핵심단어의 추출은 대량의 텍스트분석을 자동적으로 분석하는 시스템을 구축하기 위한 토대가 된다는 점에서 학문적인 기여를 하고 있다. 특히 시계열로

나타낸 오피니언 마이닝 기법을 통해 시간의 흐름에 따른 주장의 변화를 보여주는 방법론을 제시하고 있다는 점에서 공헌도가 있다. 또한 부산지역의 현안인 해수담수화 플랜트사업의 갈등구조를 선행연구와 미디어기사에서는 무엇을 핵심으로 하여 다루고 있는가에 대한 객관적인 분석결과를 제시함으로써 정책현안에 실무적인 기여를 하고 있다. 핵심단어추출의 시계열분석은 정책현안에 관한 미디어간의 사회적 토의 과정을 세밀하게 나타낸다. 정책 초기 혹은 입안의 시기에 서로 다른 인식과 논조를 가지고 기사작성을 시작한다고 하더라도, 시간이 지남에 따라 정보와 인식을 공유하고 합의하는 과정을 거치게 되어 점진적으로 차이를 좁혀가는 것이 타당한 토의의 결과라고 할 수 있다. 핵심단어추출의 시계열분석은 정책현안에 관한 미디어간의 사회적 토의 과정을 세밀하게 나타내었다. 해수담수화사업에 관한 본 분석에서는 초기부터 "원전"이라는 요소를 감지하고 시간이 지남에 따라 이해관계자간의 갈등요소들에 관한 입장의 차이를 좁혀가는 미디어 1, 2, 4, 5사의 사회적 담화과정을 파악할 수 있다. 또한 "원전"을 언급하지 않는 초기의 입장을 고수하고 있지만 질병과 사회적 갈등에 관해서는 언급을 늘려가는 미디어 3사의 변화를 파악 할 수 있었다. 이러한 결과는 해수담수화 사업에 관한 갈등을 인정하고 조절해야 한다는 사회적인 합의가 진행되고 있다는 것을 시사한다. 이러한 시사점으로부터 해수담수화시설을 통한 상수도공급 건과 관련하여 부산시민과의 갈등조절을 위한 시급한 정책적 대응이 필요하다는 정책 현안을 도출할 수 있었다.

미디어의 보도는 사회적 합의를 이끌어 내는 데, 아주 중요한 역할을 한다. 직접적인 이해당사자의 경우는 물론, 간접적인 이해당사자들은 미디어를 통하여 자신의 의견을 투영하고 의견이 다른 사람들의 의견에 대해서 이해하는 경험을 하게 된다. 만약, 미디어의 보도가 한쪽으로 편중되어 있거나 의도적으로 어떤 특정 의견만을 강조한다면, 합리적인 오피니언 형성(opinion building)을 저해할 뿐만 아니라, 사회적 합의를 통한 건전한 공공정책실현을 방해하는 요인으로 작용할 수 있다. 미디어를 통한 거시적 범위의 담화는 사회 구성원간 상호이해를 높여 인식의 차이를 줄이기 위한 프로세스를 실현하는 수단으로서의 역할을 성실히 수행해야 한다. 사회적 이슈에 대한 미디어의 거시적 담화에 있어서 성실한 토의 과정에 대한 객관적이고 투명한 평가를 위해서는 다음 세 가지 과제를 풀어야 한다.

첫째, 미시적인 단위에서의 논의에 대한 회의록 분석과 동시에, 미디어 코퍼스를 이용한 거시적인 범위에서의 논의에 대한 분석을 함께 실시하여 비교분석을 하는 것이 필요하다. 이 과정에서 미시적 코퍼스와 거시적 코퍼스에 대한 데이터베이스화 및 검색시스템을 구축하여, 객관성과 투명성을 확보하는 것이 요구된다.

둘째, 사회적 합의를 이끌어 낼 수 있도록 퍼실리테이션(facilitation) 기술을 향상시키고, 이를 지원할 수 있는 코퍼스정보통합시스템을 구축하는 것이 중요하다.

셋째, 미시적 담화와 거시적 담화를 거듭하여도 사회적 합의에 이르지 못할 경우, 어떤 시점에서 어떤 상황에서 토의를 끝맺음 하는 것이 바람직한지에 대한 마감규칙(Exiting Rules) 메뉴얼이 필요하다.

참고문헌

김동영·박제원·최재현. (2014). SNS와 뉴스기사의 감성분석과 기계학습을 이용한 주가예측 모형 비교 연구.「한국IT서비스학회지」, 13(3): 221-233.

김승현. (2016). 해수담수화플랜트 농축수의 자원화.「대한토목학회지」, 64(2): 25-29.

김원상·이종혁·박제원·최재현. (2014). 오피니언 마이닝을 통한 정당지지도 분석 기법.「한국정보기술학회논문지」, 12(10): 133-141.

김유신·김남규·정승렬. (2012). 뉴스와 주가 : 빅데이터 감성분석을 통한 지능형 투자의사결정모형.「지능정보연구」, 18(2): 143-156.

김인수·오병수. (2009). 떠오르는 물 산업 - 해수담수화.「대한토목학회지」, 57(8): 15-21.

김지훈·김건엽·남종우·김형수·장암. (2013). 세라믹 MF막의 해수담수화 전처리공정 적용성 평가.「대한토목학회 학술대회」, 2960-2961.

김충환·강두호·윤재흥. (2000). 대규모 해수담수화시설의 표준모델 적용방안 연구.「대한토목학회 학술발표회 논문집3」, 913-916.

김충환·김우구. (2000). 소규모 해수담수화시스템의 경제성분석.「대한토목학회 학술발표회 논문집3」, 921-924.

박경미·박호건·김형곤·고희동. (2011). SNS에서 오피니언마이닝 연구.「정보과학회지」, 29(11): 54-60.

박희경. (1998). 해수담수화 : 대체수자원으로서의 가능성.「대한토목학회지」, 46(4): 28-38.

박희경·박노석. (1998). 해수담수화의 국내 적용을 위한 타당성 검토.「대한토목학회지」, 46(4): 365-373.

배정환·손지은·송민. (2013). 텍스트 마이닝을 이용한 2012년 한국대선 관련 트위터 분석.「지능정보연구」,19(3): 141-156.

서지훈. (2015).「빅데이터 관리를 위한 오피니언 감성사전 모델 설계」. 박사학위논문. 인천대학교 컴퓨터공학과.

송은지. (2013). 빅 데이터를 이용한 고객평판 사례분석에 관한 연구.「한국정보통신학회논문지」, 17(10): 2439-2446.

안정국·김희웅. (2015). 집단지성을 이용한 한글 감성어 사전 구축.「지능정보연구」, 21(2): 46-67.

양민혁·정인선·김용채·조완섭. (2015). SNS 데이터를 활용한 국내대학 인식 및 선호도 분석.「한국빅데이터서비스학회 논문지」, 1(1): 1-13.

양정석·김일환. (2013). Delphi 기법을 이용한 해수담수화 플랜트 유망 국가 분석.「대한토목학회 논문집」, 33(6): 2351-2357.

우성우. (2012). 해수담수화 플랜트 시스템 설계 및 시공 기술 - 부산시 기장군 역삼투압 방식 해수담수화 플랜트 중심으로.「대한기계학회 춘추학술대회」, 125-141.

이상호·최준석·황태문. (2013). 최근 이슈가 되고 있는 차세대 해수담수화 기술 연구에 대하여 토목·환경분야가 어떤 형태로 분야발전을 이룰 수 있을까요?.「대한토목학회지」, 61(5): 102-107.

이윤주·서지훈·최진탁. (2014). SNS 텍스트 콘텐츠를 활용한 오피니언마이닝 기반의 패션 트랜드

마케팅 예측 분석. 「정보기술학회논문지」, 12(12): 163-170.
이태호. (2015). “해양심층수를 마신다고?” 부산일보. 2015.12.4.
임시영·임용민·이재용. (2014). 텍스트마이닝 기법을 이용한 U-City와 Smart City의 연구 동향에 대한 분석. 「한국지형공간정보학회지」, 22(3): 87-97.
한상진. (2016). 역량의 환경정의 관점에서 본 핵발전소 주변 주민의 건강피해와 삶의 질 문제. 「환경사회학연구 ECO」, 20(1): 283-315.
Jeong, H.Y., Shiramatsu, S., Kobayashi, K., and Hatori, T., (2008). Discourse Analysis of Public Debates using Corpus Linguistic Methodologies. Journal of Computers, 3(8): 58-68.
Chul-hyun Kim, (2009). A Taxonomical Classification of Technology-based Services: Business Model Patent Analysis. The Journal of Internet Electronic Commerce Resarch, 9(2): 167-192.
Salton, G., & McGill, J.M(Eds.), (1983). Introduction to Modern Information Retrieval, McGill-Hill.

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(napsem@gmail.com).

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).

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(jhayeong@gmail.com).

홍혜승(洪慧丞): 성균관대학교 국정전문대학원에서 행정학박사 학위를 취득하였고(조직 내 사회적 관계가 구성원의 조직행태에 미치는 영향: 리더-구성원의 교환관계와 사회적 네트워크를 중심으로, 2016), 현재 성균관대학교 리더십과 공공관리연구소에서 연구원으로 재직 중이다. 관심 연구분야는 조직행태, 리더십, 인사 및 조직관리 등이며 최근 논문으로는 “심리적 계약위반이 이직의도 및 동기부여에 미치는 영향: 직장가족주의 성향의 조절효과(2012)”, “신뢰가 정치효능감과 정치참여에 미치는 영향: 구조방정식모형을 적용하여(2015)” 등이다(blossomday@skku.edu).

〈논문접수일: 2016. 11. 7 / 심사개시일: 2016. 11. 24 / 심사완료일: 2016. 12. 19〉

Abstract

A Time Series Analysis on the Daily Media Corpus of Public Project in Local Government: Focusing on the Gijang Seawater Desalination Plant Project

Lee, Tai Hun
Hong, Soon-Goo
Jeong, Hayeong
Hong, Hye Seung

The aims of this study is to develop an automated corpus analysis methodology for objectifying a consensus dissonance among stakeholders and providing a reference to support conflict management by proposing a topic-based time series analysis of news media. To understand a consensus dissonance on the seawater desalination plant project, news articles of two local newspapers and three major newspapers which reported about the issue are collected, and a topic-based time series analysis is conducted. As a result, this study clarified a consensus dissonance about the seawater desalination plant project. This study contributes to develop an automated corpus analysis system by applying Korean natural language processing and TFIDF analysis.

Key Words: Regional Innovation, Seawater Desalination Plant, Corpus Analysis, Text Mining, Co-creation