텍스트 마이닝 기법을 활용한 동남권 신공항 신문기사 분석

(Analysis of News Regarding New Southeastern Airport Using Text Mining Techniques)

한무명초*, 김양석**, 이충권***

(Mu Moung Cho Han, Yang Sok Kim, Choong Kwon Lee)

요 약

사회적 이슈는 정책의 방향을 결정하는 중요한 요인이며, 신문은 사회적 이슈를 반영하는 중요한 채널이다. 신문기사의 텍스트를 분석하는 것은 사회적 이슈를 이해하는 데 기여할 수 있지만, 대규모의 비정형 데이터인 뉴스를 수작업으로 분석하는 것은 매우 어렵다. 따라서 본 연구는 텍스트 분석기법과 연관분석 기법을 활용해 비정형 신문기사 내용을 정형화하여 사회적 이슈의 이해관계자들 간 관점 차이를 시스템적으로 분석하는 것을 목적으로 한다. 본 연구 수행을 위해 각 지역을 대표하는 신문사(조선일보, 중앙일보, 동아일보, 매일신문, 부산일보)를 선정한 후 기사 115건과 댓글 6,772건을 2주간 수집하여 분석하였다. 연구 결과 전국 일간지들은 해당 지역과 정치적인 관계에 초점을 맞춘 반면에, 지역 일간지들은 속해 있는 지자체를 대변하는 논조로 기사가 작성된 측면이 강하게 나타났다.

■ 중심어 : 텍스트 마이닝 ; 연관분석 ; 워드 클라우드 ; 신공항

Abstract

Social issues are important factors that decide government policy and newspapers are critical channels that reflect them. Analysing news articles can contribute to understanding social issues, but it is very difficult to analyse the unstructured large volumes of news data manually. Therefore, this study aims to analyze the different views among stakeholders of a specific social issue by using text analysis, word cloud analysis and associative analysis methods, which systematically transform unstructured news data into structured one. We analyzed a total of 115 news articles and a total of 6,772 comments, collected from the selected newspapers (Chosun-Il-bo, Joongang-Il-bo, Donga-Il-bo, Maeil Newspaper, Busan-Il-bo) for two weeks. We found that there are significant differences in tone between newspapers. While nation-wide daily newspapers focus on political relations with local areas, local daily newspapers tend to write articles to represent local governments' interests.

■ keywords : text mining ; association analysis ; word cloud ; new southeastern airport

Ⅰ. 서 론

많은 인적, 물적 자원이 투입되는 대규모 국책사업은 광범위하고 지속적으로 국민 생활에 영향을 미친다. 이와 같은 사업의 추진에서 이해당사자들 간의 조정과 협력에 실패하면 사회적 갈등을 유발하고 많은 비용과 소모적인 논쟁을 불러오기도 한다.

최근 국책사업으로 추진된 동남권 신공항 건설에서 밀양을 선호한 대구·경북·울산·경남과 가덕도를 선호한 부산은 부지선정 과정에서 많은 갈등과 논쟁을 경험하였다. 정부가 동남권 신공항 건설을 백지화하고 기존의 김해공항을 확장하는 것으로 결론을 내리고 발표함으로써 지역 간의 갈등은 더욱 심각해졌으며 논란을 불러일으켰다. 따라서 언론사들은 동남권 신공항과

* 정회원, 계명대학교 경영정보학과
** 정회원, 계명대학교 경영정보학과
*** 정회원, 계명대학교 경영정보학과

접수일자 : 2017년 03월 21일
수정일자 : 2017년 03월 29일
게재확정일 : 2017년 03월 31일
교신저자 : 이충권 e-mail : cklee@kmu.ac.kr

관련하여 많은 신문기사를 게재하였고, 다양한 견해를 가진 독자들은 댓글로 자신의 의견을 표현하였다. 이에 따라 기사와 댓글들은 동남권 신공항에 대한 각 언론사의 논조와 이해 지역의 입장을 파악할 수 있는 중요한 정보를 제공한다.

기사와 댓글은 비정형 데이터인 텍스트로 존재한다. 이러한 텍스트는 현실 세계에서 정보를 표현, 전달, 교환하는 가장 대표적인 수단 중 하나이다[1]. 최근 웹과 소셜 미디어의 발달로 인해 대량의 비정형 텍스트가 기하급수적으로 증가되어 유통되고 있다. 이와 같은 다량의 비정형 텍스트에 대한 분석을 통해 의미 있는 가치를 찾고자 하는 기법으로 텍스트 마이닝은 다양한 분야에서 적용되고 있다[1,2,3,4,5].

사회적 이슈에 대해 이해관계자들은 서로 다른 입장을 가질 수 있으며, 소통 채널에 따라 다른 표현을 사용할 수 있다. 이 문제에 대한 실증적 분석을 위해 본 연구에서는 신공항 선정과 관련된 뉴스와 댓글을 분석하였다. 이를 위해 각 지역을 대표하는 언론사를 선정한 후 뉴스와 댓글을 수집하여 다음과 같은 연구를 수행하였다. 첫째, 콘텐츠 분석 기법을 활용하여 텍스트 빈도와 워드 클라우드 시각화를 수행하였다. 둘째, 연관 규칙 기법을 사용하여 사회적 이슈에 대한 연관 키워드 분석을 수행하였다. 이는 단순히 많이 출현한 단어의 빈도를 보여주기보다 특정한 주제와 관련된 키워드들이 어떻게, 어떤 수준으로 연관되어 있는지를 보여준다는 점에서 의의가 있다.

Ⅱ. 관련 연구

1. 신공항

동남권에 있는 대표적인 공항인 김해 국제공항의 사용자 수가 빠르게 증가하여 포화상태에 이를 것이라는 지적에 따라 신공항의 필요성은 노무현 정부 때 처음으로 제기되었다. 그 후 2007년 대선에서 이명박 후보의 공약에 포함되었고, 이명박 정권에서 타당성 및 입지조사를 마쳤으나 경제성이 낮아 백지화 됐다. 그 후 2012년 대선에서 유력한 후보였던 박근혜와 문재인 후보의 공약으로 다시 제시되었다. 박근혜 정부는 2015년 6월 파리공항공단 엔지니어링에 신공항 타당성 평가 연구용역을 지시했다. 그 후 부지선정 결과 발표가 가까워지면서 경남 밀양과 부산 가덕도를 선호하는 지역 간 갈등이 심해졌다. 박근혜 정부는 2016년 6월 22일 신공항 부지선정을 놓고 치열한 경쟁을 펼쳐온 경남 밀양과 부산 가덕도가 아닌 김해공항 확장으로 결론을 내렸다.

2. 텍스트 마이닝

텍스트 마이닝은 자연어 처리 기술을 기반으로 직접적인 연관을 보여주지 않는 비정형 텍스트에서 숨겨진 관계 또는 패턴을 도출하여 의미 있고 활용 가치가 높은 정보 또는 지식을 창출하는 기법이다[6,7]. 텍스트 마이닝은 분석 목적에 따라 벡터, 행렬, 계층 등의 다양한 형태로 표현될 수 있지만, 일반적으로 벡터공간모델을 사용한다[8,9]. 텍스트 마이닝을 위한 분석 방법에는 감성 분석(Sentiment Analysis), 정보 추출(Information Extraction), 네트워크 분석(Network Analysis), 텍스트 분류(Classification), 텍스트 군집화(Clustering) 등이 있다[10].

텍스트 분석에서 문서 중 단어의 중요도를 측정하는 방법은 특정한 단어가 한 문서 내에서 얼마나 자주 반복되는지, 그리고 문서 그룹 내에서 동일한 단어가 얼마나 많이 출현하는지를 측정한다. 이를 위해 단어 빈도(Term Frequency: TF)와 문서 빈도(Document Frequency: DF)를 측정한다. 단어 빈도는 특정한 단어가 문서 내에 얼마나 자주 등장하는가를 나타내는 값이다. 특정한 단어가 많이 반복될수록 문서에서 중요한 단어라고 생각할 수 있다. 문서 빈도는 문서 그룹 내에서 특정한 단어가 자주 반복되는 정도를 측정한다. 역문서 빈도(Inverse Document Frequency: IDF)는 DF 값의 역수이다. 단어가 많은 문서에서 반복적으로 나타나면, 문서를 서로 구별하는 데 사용될 가치가 감소한다. 따라서 IDF를 활용해 단어의 중요도를 나태 낼 수 있다. TF-IDF는 TF와 IDF를 곱한 값으로 문서 그룹이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한가를 나타내기 위해 사용한다. 이를 이용하여 검색 엔진에서 검색 결과의 순위를 결정하거나, 문서에서 핵심어를 추출하거나, 문서들 사이의 비슷한 정도를 측정하기 위해 사용한다[11,12,13].

3. 연관성 분석

연관성 분석은 항목 간의 상호 관계를 분석하는 것이다. 텍스트 분석에서는 단어와 단어의 상호 관계를 분석하기 위하여 단어의 동시발생(Co-Occurrence)을 분석한다. 동시발생이란 두 단어가 특정 순서로 자주 발생할 확률을 나타내는 언어학 용어이다. 동시에 발생되는 단어는 의미적 접근성 지표(An Indicator of Semantic Proximity)로 해석될 수 있으며, 이러한 두 단어는 상호 의존성을 가진다. 또한, 동시발생 분석은 언어 구조에 대한 발견과 발전이 가능하다[14].

연관성 분석의 측도는 지지도(Support), 신뢰도(Confidence)와 향상도(Lift) 값을 잘 보고 결정해 한다. 지지도란 전체 문서 중 단어 A와 단어 B가 동시에 발생하는 정도를 나타낸다. 신뢰도는 단어 A를 포함한 문서 중에서 단어 A와 단어 B가 함께 발생할 확률이 어느 정도인가를 나타낸다. 향상도는 단어 A가 발생하지 않았을 때 단어 B가 발생할 확률에 비해 단어 A가 발생하였을 때 단어 B의 발생 확률 증가 비율이다. 향상도가 1이면 두 단어의 발생 연관이 서로 관련이 없는 우연적인 결과이고, 1보다 크면 두 단어의 발생 연관이 우수하며, 1보다 작으면 두

단어의 발생 연관은 우연적 기회보다 좋지 않음을 의미한다[15].

4. 워드 클라우드

워드 클라우드(Word Cloud)는 대표적인 텍스트 시각화기법 중 하나이다. 시각화는 미적 형태와 기능성을 가지는 것으로 데이터의 연결과 그룹화를 표현하는 데 초점을 둔다. 그래프의 형태나 표로 텍스트와 같은 비정형 데이터를 시각화하는 것에는 한계가 있고 또한 직관적이라는 시각화의 효과를 기대하기 어려운 단점이 있다. 워드 클라우드는 최소의 의미를 지니는 문장 구성 성분인 형태소를 분석하고 그 빈도에 따라 문자의 크기를 결정한다. 이러한 시각화는 텍스트에서 키워드의 빈도를 직관적이고 빠르게 인지할 수 있는 장점이 있다[15,16]. SNS의 키워드 분석과 같은 다양한 부분에서 사용된다[17].

5. 래피드마이너

래피드마이너는 데이터 마이닝과 기계학습의 플랫폼으로 데이터 로딩, 전처리, 변환, 시각화, 예측분석과 통계모델링을 갖추고 있는 자바 프로그래밍언어이다[18]. 래피드마이너는 데이터 마이닝을 위한 독립형 소프트웨어로서 텍스트에서 특정 단어의 발생빈도와 키워드의 관계를 찾을 수 있다[19].

III 연구 설계

1. 자료 수집 및 설계

본 연구는 신문기사와 댓글의 텍스트 분석을 위하여 2016년 6월 28일부터 2주간 '신공항'이라는 주제어로 조선일보, 중앙일보, 동아일보, 매일신문과 부산일보에서 기사 115건과 댓글 6,772건을 수집하였다. 지방 일간지인 매일신문과 부산일보 기사의 댓글은 수가 미비하여 댓글 수집에서 제외하였다. 신문사별 분석을 위한 자료수집 결과는 표 1과 같다.

표 1. 신문사별 기사와 댓글 수

	기사 수	댓글 수
조선일보	21	3,409
중앙일보	18	1,188
동아일보	19	2,175
매일신문	47	
부산일보	10	
합 계	115	6,772

수집된 비정형 데이터를 정형 데이터로 만들기 위해 한나눔 형태소분석기를 이용해 명사, 형용사, 부사를 추출하여 CSV 파일로 변경한다. 다음으로 래피드마이너를 사용하여 Term Frequency와 Association Rules를 이용한 빈도분석과 텍스트의 연관규칙을 분석한다. 마지막으로 R 프로그램을 이용해 빈도분석 결과를 워드 클라우드로 시각화했다. 기사와 댓글 분석 과정은 그림 1과 같다.

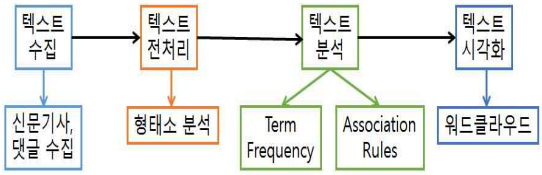

그림 1. 텍스트 마이닝 과정

IV. 연구방법

1. 텍스트 전처리

명사는 개체에 대한 속성과 감성어휘를 많이 나타내고, 형용사와 동사는 주체자의 주관적인 의견이나 존재 그리고 평가를 위한 정보를 내포하고 있으며, 부사는 다양한 표현방법과 수식어로 문서에 분포된다[20]. 따라서 한나눔 형태소분석기를 이용해 수집한 데이터에서 형용사, 명사, 부사로 추출하였다.

2. 텍스트 분석

기사와 댓글의 빈도분석은 그림 2와 같이 래피드마이너의 Process Documents From Data 오퍼레이터를 이용해 문자열 속성 벡터를 생성했다. 이 오퍼레이터의 Vector Creation 파라미터는 Binary Term Occurrences로 설정해 토큰 된 단어의 빈도분석이 가능한 0과 1의 매트릭스를 생성한다. Process Documents From Data 오퍼레이터의 내부 처리 과정은 다음과 같다. 첫째, Tokenize 오퍼레이터를 이용해 기본 파라미터 값인 Non Letters 단위로 토큰 한다. 둘째, Filter Tokens(by length) 오퍼레이터를 이용하여 최소 문자 2와 최대 문자 10으로 설정해 단어 길이가 2글자 이상 10글자 이하인 단어를 추출한다. 셋째, Generate N-Gram(terms) 오퍼레이터는 N개의 연속적인 토큰을 연결하여 의미를 만들어 낸다. 본 연구에서는 Max Length 파라미터 값을 2로 설정했다. 넷째, Filter Stopwords(dictionary) 오퍼레이터를 이용해 추출하지 않을 단어를 파일로 정의해준다. 본 연구에서는 기사의 빈도분석에만 Stopwords를 사용하였으며, '신공항', '공항'으로 정의하였다. 신공항 주제로 기사를 분석하고 있으므로 두 단어는 많은 빈도를 보이나 중요도는 낮을 것으로 판단되기 때문이다.

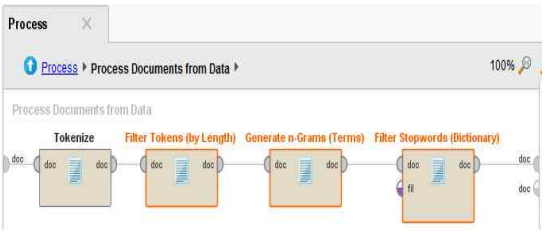

그림 2. Process Documents From Data 과정

텍스트의 연관분석은 FT-Growth 오퍼레이터와 Create Association Rules 오퍼레이터를 이용했다. FT-Growth 오퍼레이터는 Binomial 값만을 계산하므로 Numerical to Binominal 오퍼레이터를 사용해 Numeric 속성을 Binominal 값으로 변경했다. 기사 분석의 경우는 FT-Growth 오퍼레이터의 Min Support 파라미터 값을 .95로 설정했고, 댓글 분석에서는 Min Support 파라미터 값을 .01로 설정했다. 댓글이 많아지면서 단어의 종류가 다양해져 그 조합이 많아지므로 Min Support 값을 조정하였다. 텍스트 분석과정은 그림 3과 같다.

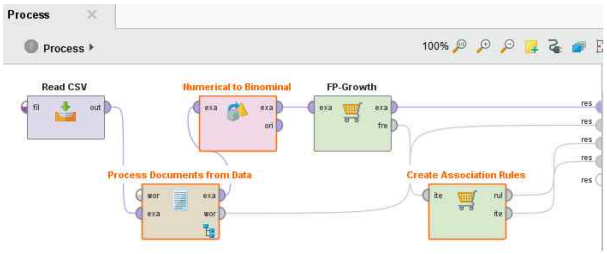

그림 3. 텍스트분석 과정

3. 텍스트 시각화

각 언론사의 신문기사 빈도는 10 이상, 댓글 빈도는 5 이상인 단어를 워드 클라우드로 시각화하였다. 댓글의 특성상 글의 길이가 짧아 기사의 빈도보다 상대적으로 작게 5 이상으로 정하였다. 워드 클라우드는 R프로그램에서 Word Cloud 패키지를 이용하여 작성하였다.

V. 연구 결과

1. 신문사별 빈도분석

신문사별 기사 빈도분석 결과는 표 2와 같다. 보수 계열을 대표하는 전국 일간지인 조선일보의 기사 빈도분석 결과는 '대구', '지역', '이전' 순으로 나타났으며, '대통령', '의원', '정부'라는 정치와 관련된 단어도 높은 순위를 보인다. 중앙일보의 기사 빈도분석 결과는 '대구', '정부', '이전' 순으로 분포되었으며, '정부', '의원', '국민'이라는 정치와 관련된 단어도 높은 순위를 보였다. 동아일보 기사의 빈도분석 결과는 '대통령', '의원', '이전' 순으로 분포되었으며, 이러한 단어도 정치와 연관된 단어를 포함하고 있었다. 다른 중앙일간지와 다르게 동아일보는 '문제', '특사'라는 단어가 눈에 띈다. 조선일보와 중앙일보는 '사드'라는 단어가 일곱 번째, 여섯 번째로 나타나고 있다. 이 단어는 또 다른 지역갈등을 촉발할 가능성이 높은 단어이다.

지역 일간지인 매일신문의 기사 빈도분석 결과는 '대구', '백지화'라는 단어로 대구의 입장을 나타내고 있으며, 부산일보의 기사 빈도분석 결과는 '부산', '결정'이라는 단어로 부산의 입장을 나타내고 있다. 이해당사 지역의 일간지들은 각 지역의 명칭이 높은 순위를 나타내고 있다.

정기선(2005)은 지역갈등의 인식변화에 관한 연구에서 1988년에는 지역갈등의 원인을 경제발전정책 탓으로 보았고, 2003년에는 지역주민의 의식과 정치인의 선거운동을 주요 원인으로 보았다 [21]. 대부분 기사에서 알 수 있듯이 신공항 부지선정에서도 정치와 관련된 논조가 지배적인 것으로 나타났다.

표 2. 신문사별 기사 빈도분석 결과

조선일보		중앙일보		동아일보		매일신문		부산일보	
대구	144	대구	54	대통령	82	대구	356	지역	51
지역	77	정부	40	의원	48	이전	311	부산	40
이전	76	이전	36	이전	42	정부	163	김해공항	37
대통령	67	의원	33	대구	36	대구_공항	148	정부	27
의원	64	지역	32	김해공항	30	의원	124	배치	25
정부	61	사드	29	문제	29	지역	103	사드	25
사드	49	배치	28	사회	28	김해공항	96	확장	24
배치	48	국토	20	지역	28	영남권	96	공사	22
결정	35	국민	17	영남권	27	확장	89	대구	19
경북	35	대표	17	특사	27	백지화	78	결정	18

전국 일간지의 댓글 빈도분석에서는 모두 '사드배치'라는 단어가 높은 순위에 나타났다. 이는 기사를 읽은 시민들은 새로운 지역갈등과 정치적 마찰을 불러올 수 있는 사드배치에 관심이 높다는 것을 알 수 있다. 조선일보는 '대구시민'이라는 단어가 높은 순위에 나타났다. 이 결과는 대구시민과 이해관계가 있는 사람이 조선일보 기사를 많이 읽었을 가능성이 높다는 것을 미루어 짐작할 수 있다. 중앙일보는 '이제와서'라는 단어가 눈에 띈다. 마지막으로 동아일보는 '특별사면'이라는 단어가 가장 높은 순위를 보인다. 이는 동아일보 기사에서 '대통령'이 가장 높은 빈도를 보이는 것과 연관된 것으로 보인다. 전국 일간지 댓글 빈도분석 결과는 표 3과 같다.

표 3. 전국 일간지 댓글 빈도분석 결과

조선일보		중앙일보		동아일보	
사드배치	108	사드배치	54	특별사면	46
우리나라	66	새누리당	13	국민통합	37
새누리당	31	우리나라	11	공군기지	30
대구시민	25	다른지역	9	김해공항	25
김해공항	23	이제와서	9	사드배치	20

워드 클라우드는 조선일보와 중앙일보의 기사 빈도 중에서는 '대구'가 가장 크게 보이고, 두 언론사의 댓글에서는 '사드배치'가 가장 크게 시각화되어 두 단어를 직관적으로 연결해 볼 수 있다. 중앙일보의 댓글에서는 '특별사면'이라는 단어의 빈도가 높게 나타났고, 그 이유는 중앙일보 기사에서 '대통령', '특사'라는 단어를 통해 그 관련성을 직관적으로 알 수 있다. 이해당사 지역의 일간지 들은 각 지역 명칭을 크게 보여 줌으로써 지역을 대변하고 있음을 알 수 있다. 워드 클라우드 내용은 그림 4에서부터 그림 7까지와 같다.

그림 4. 조선일보 기사 & 댓글

그림 5. 동아일보 기사 & 댓글

그림 6. 매일신문 기사 & 부산일보 기사

그림 7. 중앙일보 기사 & 댓글

2. 신문사별 연관분석

신공항에 대한 신문사별 논조를 파악하기 위해 '신공항'이라는 단어와 동시발생이 많은 단어 중 지지도가 가장 높은 단어는 표4와 같다. 조선일보는 6개의 규칙을 통해 '지역', '이전', '대구', '정부', '공항'이라는 단어가 연관되어 있다. 중앙일보는 6개의 규칙을 통해 조선일보와 같은 결과를 보인다. 다만 지지도와 신뢰도에서 조금의 차이를 보인다. 동아일보는 2개의 규칙을 통해 '대구', '정부', '발표'라는 단어가 연관되어 있다. 중앙일간지는 모두 '신공항'이라는 단어와 연관된 단어가 매우 비슷하다. 이는 중앙일간지는 모두 신공항을 지역과 정치적인 관계에 초점을 맞추고 있음을 알 수 있다. 매일신문은 2개의 규칙을 통해 '영남권', '백지화', '영남권_신공항'이라는 단어와 연관되어 있다. 이는 대구의 입장인 신공항 백지화를 주장하고 있다. 부산일보는 2개의 규칙을 통해 '확장', '김해공항', '김해공항_확장'이라는 단어가 연관되어 있다. 이는 부산의 입장에서 신공항이 부지 선정이 아니라 확장으로 결정되었음을 전달한다. 이와 같은 내용의 그래프는 그림 8에서부터 그림 12까지와 같다.

표 4. 신문사별 기사 연관분석 결과

	신공항
조선일보	지역, 이전, 대구, 정부, 공항
중앙일보	지역, 이전, 대구, 정부, 공항
동아일보	대구, 정부, 발표
매일신문	영남권, 백지화, 영남권_신공항
부산일보	확장, 김해공항, 김해공항_확장

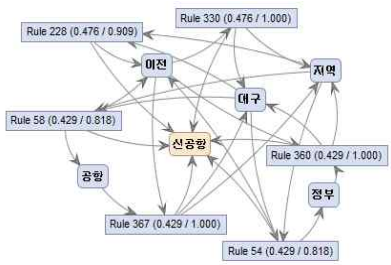

그림 8. 조선일보 기사 연관분석

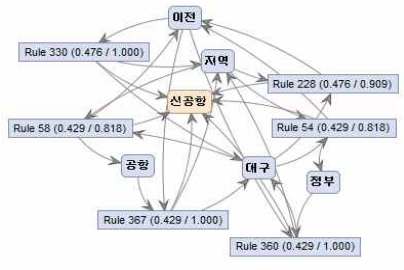

그림 9. 중앙일보 기사 연관분석

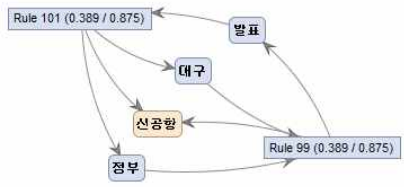

그림 10. 동아일보 기사 연관분석

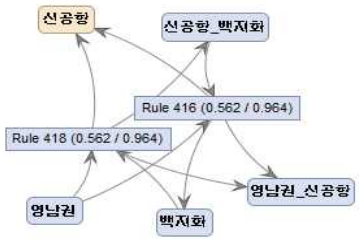

그림 11. 매일신문 기사 연관분석

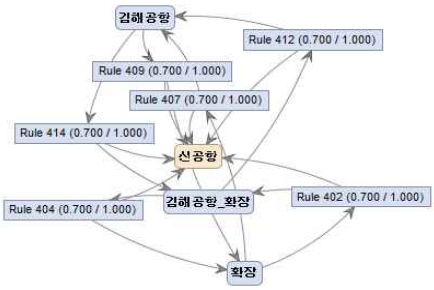

그림 12. 부산일보 기사 연관분석

6. 결론

본 연구는 비정형 데이터인 신문기사 내용을 정형화해 서로 다른 이해당사자들의 관심 표현의 차이를 세 가지 유형(전국일간지, 대구 지방지, 부산 지방지)의 신문사별 기사 내용과 독자들의 댓글을 바탕으로 분석하였다. 이를 위해 신공항 부지선정 결과와 관련해 전국 일간지인 조선일보, 중앙일보, 동아일보와 이해당사 지역 일간지인 매일신문과 부산일보의 기사와 댓글의 콘텐츠 분석과 연관분석을 하고, 분석 결과의 시각화를 제시하였다.

이를 통해 전국 일간지의 빈도분석 결과는 이해당사 지역인 '대구', '지역', '의원'이라는 단어가 10위권에 공통으로 나타났다. 특히 조선일보와 중앙일보는 '정부', '사드', '배치'라는 단어까지 공통적으로 나타났으며 연관분석 결과 또한 같다. 이는 신공항을 바라보는 논조가 매우 비슷함을 알 수 있다. 또한, 이 기사의 독자들의 반응 또한 '사드배치'라는 단어로 같게 나타나고 있다. 지역 일간지는 속해 있는 지자체 명이 상위에 나타났으며, 매일신문은 '백지화'랑 부산일보는 '김해공항_확장'이라는 단어로 지자체를 대변하는 논조로 기사가 작성되었다. 전국 일간지의 댓글 분석에서는 또 다른 지역 관심사이며 정치적 분쟁을 불러일으킬 수 있는 사드배치에 높은 관심을 보인다. 따라서 추후 사드배치 문제는 많은 분란을 일으킬 것으로 예상된다. 보수 성향을 나타내는 전국 일간지 3곳은 전체적으로 비슷한 관점으로 기사를 생성하였고, 그 기사를 읽고 댓글을 작성한 시민들 또한 비슷한 여론을 형성하였다. 이는 본 연구에서 단어 빈도와 연관분석의 시각화를 통해 객관화된 데이터로 확인할 수 있었다.

본 연구는 각 신문사의 논조와 독자들의 의견을 객관적인 데이터로 보여주기 위해 빈도분석, 연관분석, 시각화를 사용했다는 점에서 학문적 기여를 하였다. 그러나 본 연구는 단어 수준의 분석만을 수행하였다는 한계가 있다. 좀 더 개선된 분석을 위해서는 단어뿐만 아니라 의미적 차원의 분석이 필요할 것이다.

References

[1] I.H. Witten, "*Text Mining, Practical Handbook of Internet Computing,*" CRC Press. 2004.

[2] M.A. Hearst, "*Untangling Text Data Mining*", Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. 1999.

[3] R.J. Mooney and R. Bunescu, "Mining Knowledge from Text using Information Extraction," *ACM SIGKDD Exploration Newletter*, vol. 7, no. 1, pp. 3-10, Jun. 2005.

[4] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1-47, 2002.

[5] F. Sebastiani, "Classification of Text, Automatic," *The Encyclopedia of Language and Linguistics*, vol. 14, pp. 457-462, 2006.

[6] P. Judita, M. Stevenson, and R. Gaizauskas, "Exploring relation types for literature-based discovery," *Journal of the American Medical Informatics Association*, ocv002, pp. 987-992,

May. 2015.
[7] F. Ronen and I. Dagan, "Knowledge Discovery in Textual Databases (KDT)," *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, vol. 95, pp. 112-117, 1995.
[8] S. Gerard, A. Wong, and C.S. Yang, "*A Vector Space Model for Automatic Indexing,*" *Communications of the ACM*, vol. 18, no. 11, pp. 613-620, Nov. 1975.
[9] S. Anna, P. Andritsos, and N. Nicoloyannis, "*Overview and Semantic Issues of Text Mining,*" *ACM SigMod Record*, vol. 36, No.3, pp. 23-34, Sep. 2007.
[10] W. Fan, L. Wallace, S. Rich, & Z. Zhang, Tapping the power of text mining. *Communications of the ACM*, vol. 49, no. 9, pp. 76-82, 2006.
[11] https://ko.wikipedia.org/wiki/TF-IDF 2016. 9. 19. 검색
[12] H. Jiawei, J. Pei, and M. Kamber, "*Data mining: Concepts and Techniques*," 3rd Edition, Morgan Kaufmann Publishers, 2011.
[13] J.H. Park and S. Min, "A Study on The Research Trends in Library & Information Science in Korea Using Topic Modeling," *Journal of the Korean Society for information Management*, vol. 30, no. 1, pp. 7-32, 2013.
[14] R. Paul and Kroeger, "*Analyzing Grammar: An Introduction*," Cambridge University Press, 2005.
[15] 서강수, "*데이터 분석 전문가 가이드*", 한국데이터베이스진흥원, 2014.
[16] 노형남, "워드 클라우드에 의한 환대 경영 전략," *관광연구*, 제29권, 제4호, pp. 335-354, 2014.
[17] T. Hammond, T. Hannay, B. Lund, and J. Scott, Social bookmarking tools (I), *A general review. D-Lib Magazine,* vol. 11, no. 4, 2005.
[18] P. Abhinn, "Study and Analysis of K-Means Clustering Algorithm Using Rapidminer," *International Journal of Engineering Research and Applications*, vol. 1, no. 4, pp. 60-64, Dec. 2014.
[19] A. Kumar, P. Thakur, K. Gupta, and A. Pal, "*Text mining approach to analyse the relation between obesity and breast cancer data,*" *International Letters of Natural Sciences*, vol. 44, no. 1, pp. 1-9, 2015.
[20] 강대국, 박용태, "리뷰 기반의 모바일 서비스 고객 요구사항 특성 분석," *한국경영과학회 추계학술대회, 방위사업청 무기체계 시험평가 세미나 논문집*, pp. 945-951, 2012.
[21] 정기선, "지역감정과 지역갈등인식의 변화 1988년과 2003년 비교," *한국사회학*, 제39권, 제2호, pp. 69-99, 2005.

저 자 소 개

한무명초(정회원)

2006년 방송통신대학교 컴퓨터과학학과 학사 졸업.
2009년 계명대학교 전산교육학과 석사 졸업.
2016년 계명대학교 경영정보학과 박사 졸업.
<주관심분야 : 데이터마이닝, 정보기술, 지식관리>

김양석(정회원)

1995년 서울시립대학교 경제학과 학사 졸업.
2004년 University of Tasmania 컴퓨터 공학 석사 졸업
2009년 University of Tasmania 컴퓨터 공학 박사 졸업.

<주관심분야 : Big Data, Data Analytics, Knowledge-based System>

이충권(정회원)

1995년 계명대학교 경영정보학과 학사 졸업.
1999년 Southeast Missouri State University MBA 졸업
2003년 University of Nebraska-Lincoln 박사 졸업.
2003-2006년 Georgia Southern University 조교수
<주관심분야 : Big Data, Text Mining, IT Jobs>