Research Paper Informatization Policy Vol. 27, No.4 (2020), pp.024-046
https://doi.org/10.22693/NIAIP.2020.27.4.024

키워드 네트워크 분석을 이용한 공공데이터 수요 예측

이 재 원*

요약 본 연구는 키워드 네트워크 분석을 이용하여 공공데이터 수요(즉, 공공데이터 제공신청, 검색 질의 등)를 적시에 예측하는 방법을 제안한다. 분석 결과에 따르면, 수요가 높은 토픽에 속하는 공공데이터는 대부분 국내 공공데이터 포털(data.go.kr)에서 제공되고 있지만, 토픽 연관 분석을 통해 예측된 이용자의 실제 요구와 관련된 공공데이터는 거의 제공되지 않고 있다. 공공데이터를 제공(또는 선정)할 때, 이용자의 공공데이터 제공신청과의 관련성보다 공공데이터 토픽과의 관련성이 우선시되기 때문이다. 제안된 키워드 네트워크 분석 프레임워크는 실제 공공데이터 제공신청을 바탕으로 이용자들의 수요를 빠르고 쉽게 예측할 수 있으므로, 향후 공공기관(중앙부처·지방자치단체·산하기관)의 공공데이터 정책 수립에 이바지할 수 있을 것으로 기대된다.

주제어 : 공공데이터, 제공신청, 키워드 네트워크 분석, 수요 예측, 개방 정책

Forecasting Open Government Data Demand Using Keyword Network Analysis

Lee, Jae-won*

Abstract This study proposes a way to timely forecast open government data (OGD) demand(i.e., OGD requests, search queries, etc.) by using keyword network analysis. According to the analysis results, most of the OGD belonging to the high-demand topics are provided by the domestic OGD portal(data.go.kr), while the OGD related to users' actual needs predicted through topic association analysis are rarely provided. This is because, when providing(or selecting) OGD, relevance to OGD topics takes precedence over relevance to users' OGD requests. The proposed keyword network analysis framework is expected to contribute to the establishment of OGD policies for public institutions in the future as it can quickly and easily forecast users' demand based on actual OGD requests.

Keywords : open government data(OGD), offer request, keyword network analysis, demand forecasting, OGD policy

Received Aug 29, 2020; Revised Sep 22, 2020; Accepted Sep 24, 2020
* Executive Principal, National Information Society Agency(jaewonlee@nia.or.kr)

Ⅰ. 서론

1. 연구의 필요성 및 목적

사물인터넷(IoT), 빅데이터 분석, 인공지능(AI) 등 다양한 ICT 기술의 발전에 따라, 전 세계적으로 데이터를 중심으로 다양한 형태의 경제적·사회적 가치가 창출되고 있다. 최근 모바일 기기의 보편화로 인해 다양한 모바일 앱 서비스가 개발되고 있는데, 이들은 공공기관(중앙부처·지방자치단체·산하기관 등)에서 생산하는 공공데이터를 활용함으로써 자체 서비스의 품질에 대한 높은 신뢰를 확보하고 있다. 예를 들어, 모바일 기기에서 제공되는 날씨 앱이나 버스 위치 앱 등은 공공기관에서 생산하는 데이터 즉, 공공데이터를 활용함으로써 이용자들의 날씨 및 버스 위치 정보의 정확도에 대한 신뢰도를 높이고 있다. 또한, 최근 국내에서는 코로나19라는 국가적 위기 상황에서 공공기관에서 보유하고 있는 공적 마스크 정보를 제공하고, 민간 기업이 국민에게 필요한 마스크 정보 앱(마스크 보유현황, 약국 위치 등)을 신속하게 만들어 제공하였다. 공공기관에서 생산한 공공데이터를 민간 기업이 이용하여 코로나19에 대응한 K-방역은 공공데이터 생산 및 소비 활동을 보여주는 대표적인 국내 성공 사례이다. 우리나라뿐만 아니라, 미국·영국 등 주요 선진국들도 공공데이터를 개방함으로써, 새로운 경제적·사회적 가치를 창출해오고 있다. 특히, 공공데이터를 지속적·체계적으로 국민에게 제공하기 위해 세계 각국은 자국의 공공데이터 개방 포털을 구축하여 운영해 오고 있다.

우리나라도 국민의 공공데이터 대한 이용권을 보장하고, 공공데이터의 민간 활용을 통한 삶의 질 향상과 국민경제 발전에 이바지하기 위해 2013년부터 「공공데이터의 제공 및 이용 활성화에 관한 법률(약칭: 공공데이터법)」을 시행하고 있다. 이에 따라, 공공기관이 생성·보유 중인 공공데이터는 공공데이터 포털(data.go.kr)을 통해 국민, 기업 등 민간 이용자들에게 개방되고 있다. 공공데이터 포털을 통해 공공기관에서 보유 중인 공공데이터를 파일이나 오픈 API 등 다양한 형태로 통합 제공하는 것은 민간 이용자들의 공공데이터 접근성을 크게 높일 수 있다. 공공데이터 전략위원회 기본계획(Open Data Strategy Council, 2019)에 따르면, 최근 3년간 공공데이터 개방은 약 53% 증가하여 32,743개를 개방 중이며, 공공데이터에 대한 민간 이용은 약 656% 증가한 1,235만 건에 달한다(19년 기준). 이와 같은 노력의 결과, 우리나라는 OECD 공공데이터 평가에서 연속 3회(15년, 17년, 19년) 세계 1위를 달성하였고, ODB(Open Data Barometer) 평가 순위도 상승하는 추세이다. 그러나, 이러한 정부의 다각적 노력에도 불구하고, 공공데이터 관련 개방 정책이나 서비스들이 이용자 수요를 제대로 반영하지 못하여 쓸만한 데이터 없다는 비판이 꾸준히 제기되고 있다(Seo & Myeong, 2014; National Information Society Agency, 2017; Lee & Park, 2019).

이용자들이 원하는 공공데이터를 적시에 개방하기 위해서는 공공데이터를 실제 필요로 하는 이용자들의 수요를 정확히 파악하는 것이 중요하다. 일반적으로, 공공데이터 개방 정책 수립에 필요한 실제 공공데이터 이용자들의 수요를 파악하는 방법은 크게 2가지로 나눌 수 있다. 첫 번째 방법은 사례 연구(Dawes, 2016; Ruijer, 2020), 웹사이트 분석(Lourenço, 2015; Wang, 2020), 설문 연구(Worthy, 2015), 문헌 연구(Ohemeng & Ofosu-Adarkwa, 2015; Han, et al., 2020) 등과 같이 제한된 자료 수집 및 연구 환경을 설정하여 분석하는 것이다. 이들은 특정 사례(혹은 기관)의 특성을 반영하여 공공데이터 수요에 대한 분석을 수행할 수 있다는 장점이 있으나, 이용자들이 원하는 데이터를 파악하기까지 상당한 비용과 시간이 소요되는 단점이 있다. 특히, 코로나19와 같이 긴급 상황이 발생하였을 때, 짧은 기간 안에 모든 이용자의 공공데이터 수요를 정확히 파악하기에는 적절하지 않다. 또한, 선행 연구에서 수행한 연구 방법들은 제한된 대상(분류, 도메인, 참여자 등)으로부터 결론이 도출되었기 때문에, 공공데이터 개방 정책을 수립하는데 실제 적용을

할 수 있는지는 미지수이다. 이와 같은 선행 연구의 문제점을 해결하기 위해서는 실제로 공공데이터 포털을 통해 접수된 민간 이용자들의 실제 공공데이터 수요 정보를 대상으로 분석할 필요가 있다.

두 번째 방법은 공공데이터 포털에서 운영 중인 '데이터 제공신청'에 접수된 이용자들의 실제 공공데이터 수요 정보를 분석하는 것이다. 공공데이터 제공신청 서비스는 공공데이터법 제27조(공표 제공대상 외의 공공데이터 제공신청 등)에 따라 운영되는 제도이다. 이를 통해, 민간 이용자들에게 필요한 공공데이터가 공공데이터 포털에 아직 미개방 중인 경우, 언제든지 온·오프라인 창구를 통해서 공공데이터 제공을 요청할 수 있다. 하지만, 공공데이터 포털에 접수된 공공데이터 제공신청 정보를 분석하여 이용자들의 실제 수요를 분석한 선행 연구는 많지 않다. Cho and Ha(2020)는 공공데이터 제공신청 정보를 대상으로 구조적 토픽 모델링(Structural Topic Modelling) 기법을 적용하여 공공데이터의 잠재적 토픽(Topic)을 도출하고, 토픽별 공공데이터 제공신청 변화(즉, 공공데이터 수요 변화)를 분석하였다. 토픽 모델링은 문서를 구성하는 단어의 분포를 이용하여 개별 문서를 구성하는 잠재적 토픽 분포를 추정할 수 있다(Blei & Lafferty, 2007; Suh & Shin, 2017; Cho, et al., 2018). 토픽을 중심으로 공공데이터 수요를 분석하는 것은 거시적 관점의 공공데이터 수요 패턴을 분석한 방법이다. Cho and Ha(2020)가 제안한 공공데이터 수요 분석 방법은 거시적 관점에서 공공데이터 정책을 수립할 때 유용할 수 있으나, 공공데이터 이용자들이 실제 원하는 데이터를 제공하지 못할 수 있다. 본 연구의 분석 결과에 따르면, 이용자들의 관심이 높은 일부 토픽의 경우, 공공데이터 포털을 통해 많은 데이터가 이미 제공되고 있으나 이용자들이 실제 원하는 정보는 여전히 미제공 중인 경우가 많다.

이러한 문제점을 해결하기 위해서는 토픽보다 입상(Granularity)이 작은 키워드 레벨에서 공공데이터 수요 분석을 수행하는 것이 적절하다. 본 연구는 키워드 네트워크 분석 기법을 적용하여, 키워드 레벨의 관계 정보 분석을 통해 공공데이터 수요 및 이용 상황에 따른 공공데이터 수요의 변화 추이에 대하여 분석한다. 키워드 네트워크 분석은 특정 주제 영역의 문서 집합으로부터 키워드를 추출하고, 각 키워드 쌍(Pair)의 동시 출현 빈도를 기반으로 키워드 간의 유사도를 계산하여 구성된 키워드 네트워크에서 상호 관계를 분석하는 방법이다(Iem, et al., 2015). 키워드 네트워크 분석 방법은 시간의 경과에 따라 키워드 간의 관계 변화를 파악하는 것이 가능하므로, 다양한 분야의 선행 연구에서 추세 예측 및 동향 분석을 위해 널리 이용되고 있다(Choi, et al., 2011; Park, et al., 2018; Rha, 2020). 하지만, 키워드 동시 출현성 분석을 통해서는 본 연구의 목적인 공공데이터 수요를 정확히 알아내는 것이 쉽지 않으므로, 제안된 키워드 네트워크 분석 프레임워크는 키워드 동시 출현성(Co-occurrence) 분석 및 토픽 연관(Association) 분석을 순차적으로 수행한다.

본 연구는 대국민 서비스를 제공 중인 공공데이터 포털에서 보유하고 있는 대용량 텍스트 데이터를 대상으로, 키워드 레벨에서 이용 상황별 공공데이터 수요 분석을 수행한 첫 연구라는 점에서 큰 의의가 있다. 또한, 공공데이터 이용 상황(평상, 긴급 상황 등) 구분하여 공공데이터 수요 분석을 비교한 국내·외 연구 사례는 아직 없다. 본 연구에서 제안하는 키워드 네트워크 기반 분석 프레임워크는 긴급한 상황에서도 빠르고 정확한 수요 분석을 수행할 수 있으므로, 정부에서 추진 중인 데이터 기반 과학적 행정 구현과 관련한 다양한 정책 수행의 기초 연구가 될 것으로 기대한다.

2. 연구의 방법 및 구성

본 연구는 크게 5단계로 구성된다.

첫째, 공공데이터 포털의 공공데이터 제공신청 정보를 수집한다.

둘째, 수집된 공공데이터 제공신청을 대상으로 국문을 영문으로 자동 변환, 불용어 제거 등 전처리

(Pre-processing)를 수행한다. 본 연구에서 수행한 키워드 분석을 위한 전처리 과정 및 고려사항에 대하여 상세히 설명한다. 일반적으로, 국문으로부터 키워드를 추출하는 것은 한국어 형태소 분석기에 대한 의존성이 높아지므로, 어떤 형태소 분석기를 이용하느냐에 따라 연구 결과가 상이할 수 있다. 연구의 객관성 및 수요 분석의 정확성을 확보하기 위해, 본 연구는 국문을 영문으로 번역하여 국문의 특성(띄어쓰기 오류, 조사 등)을 제거하였다.

셋째, 키워드 네트워크 분석을 위해, 공공데이터 제공신청 정보를 분석하여 주요 키워드 추출 및 키워드 간의 관계 정보 등을 분석한다. 또한, 공공데이터 이용 상황에 따른 공공데이터 수요 변화 추이에 대하여 분석한다.

넷째, 키워드 네트워크 분석을 통해 파악된 수요 정보와 공공데이터 포털에서 개방 중인 공공데이터 현황을 비교 분석하여, 향후 공공데이터 개방 방향성을 제시한다.

마지막으로, 결론 부분에 본 연구의 의의와 한계점을 기술하고, 향후 연구 과제를 제시한다.

Ⅱ. 선행 연구

1. 공공데이터 분야 연구 방법론

미국, 영국 등 해외 주요국을 중심으로 공공데이터가 국민 생활 향상과 경제 활성화를 이룰 수 있는 공공재로 인식됨에 따라, 자국의 공공데이터 개방을 확대하고 있다. 이에 따라, 공공데이터 분석을 위한 다양한 실증 연구 방법론이 제시되었다. 사례 연구(Dawes, 2016; Ruijer, 2020), 웹사이트 분석(Lourenço, 2015; Wang, 2020), 설문 연구(Worthy, 2015), 문헌 연구(Ohemeng & Ofosu-Adarkwa, 2015; Han, et al., 2020), 설계 연구(Zeleti, et al., 2016) 등이 대표적인 사례이다. 이러한 다양한 실증 연구 방법론들은 나름의 장·단점을 가지고 있다. 우선, 사례 연구는 여러 변수 사이의 상호 작용을 연구하기에 매우 유용하다. 공공데이터 포털과 같은 웹사이트 분석은 실제로 데이터 개방 구조 등에 대하여 시사점을 도출할 때 유용하다. 설문 연구는 공공데이터 생산자와 소비자의 공공데이터 개방에 대한 인식 차이(예를 들어, 공공데이터 개방 우선순위 선정에 대한 인식 차이 등)를 파악하기 위해 적용될 수 있는 연구 방법이다. 문헌 연구는 정부에서 발간하는 공식적인 문서를 중심으로 분석함에 따라, 공공데이터에 대한 정부의 방향성(의도 등)을 분석하기에 적합하다. 마지막으로, 설계연구는 공공데이터 서비스를 제공하기 위한 전체 아키텍처를 설계하기 위한 기초적인 요소를 제공하고 있다. 하지만, 이러한 연구 방법들은 제한된 자료 수집 및 연구 환경에서 결론이 도출되었기 때문에, 연구 결과를 범정부 차원에서 공공데이터 개방 정책을 수립하는데 실제 적용 가능한 방법론이 아니라는 한계점이 존재한다.

2. 공공데이터 개방 제도 분석

공공데이터 분야의 선행 연구들에 따르면, 공공데이터 개방은 정부 정책의 투명성 및 책임성 증대, 신규 서비스 개발을 통한 경제 성장 등 다양한 측면에서 사회 혁신을 촉발할 것으로 기대하고 있다(Jassen, et al., 2012; Tammisto & Lindman, 2012; Lourenço, 2015; Ruijer, 2017). 경제협력개발기구(OECD)에서 2019년 발표한 "2019년 OECD 정부 백서(Government at a Glance 2019)"는 각국의 공공데이터 개방 노력을 OUR Index(Seo, 2017)를 이용하여 데이터 가용성(Data Aailability), 데이터 접근성(Data Accessibility) 및 데이터 활용을 위한 정부 지원(government support for data re-use) 측면에서 분석하였다. 〈표 1〉은 "2019년 OECD 정부 백서"를 요약한 것으로, 한국이 해외 주요국 대비 공공데이터 개방 정책이 적극적으로 추진되고 있음을 알 수 있다. Cho and Ha(2020)는 주요국의 공공데이터 개방 정책을 정리하여 제시하고 있으며, 이를 요약하면 〈표 2〉와 같다.

〈표 1〉 OECD OUR Data Index 2019의 상위 5개국
〈Table 1〉 Top-5 countries in OECD OUR Data Index 2019

Rank	Final Rank	Data availability	Data accessibility	Government support
1	Republic of Korea	Republic of Korea	Austria	Republic of Korea
2	France	France	France	France
3	Ireland	Canada	Republic of Korea	Ireland
4	Japan	Japan	Norway	Japan
5	Canada	Mexico	Portugal	Austrailia

〈표 2〉 공공데이터 개방 관련 주요 정책
〈Table 2〉 Key policies related to OGD provision

Country	Key Policies
France	It provides high-value datasets in specific sector, improves open data portal, and promotes data utilization (such as AI Lab and digital public service incubator installation).
Ireland	It promotes the opening of high-value data, builds economic value through free re-use, and encourages citizen participation in various communities.
Canada	It reflects citizen opinions (such as UI/UX improvement of the OGD portal), and establishes the detailed action plans including milestones.
Australia	It establishes the detailed action plans with eight commitments including improved public sector data sharing and re-use and so on.

source: Cho & Ha (2020)

3. 공공데이터 개방 및 수요 분석

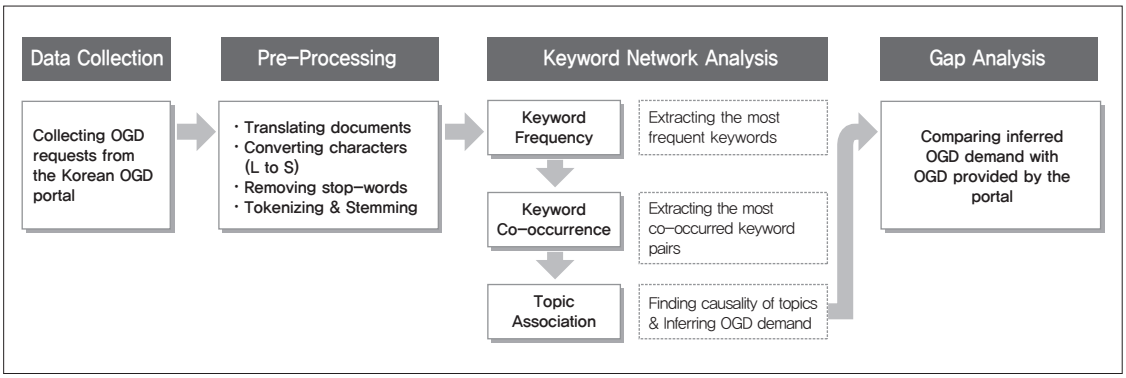
최근에 발표된 공공데이터 개방 관련 국외 연구들은 공공데이터 자체에 대한 개방·수요 분석보다 각 정부에서 운영 중인 공공데이터 포털을 중심으로 공공데이터 개방성을 측정하는 방법을 제안하였다. Thorby, et al.(2017)은 미국의 30개 도시에서 운영 중인 공공데이터 포털에 대하여 정량적, 정성적 개방성을 측정하기 위해 ODPI(Open Data Portal Index)와 DCI(Data Content Index)를 제안하였다. Chatfield and Reddick(2017)은 호주의 공공데이터 포털에 대하여 공공데이터 개방 정책 채택, 공공데이터 개수, 기계 가독성이 높은 데이터 비율, 이용자 참여 서비스(분석 도구, 해커톤 등) 제공 측면에서 분석을 수행하였다. Kubler, et al.(2018)은 메타데이터에 대한 품질을 기반으로 공공데이터 포털을 평가하기 위해 Veljković, et al.(2014)에 의해 제안된 eGovOI 모델을 이용하였다. 공공데이터 포털을 중심으로 공공데이터 개방을 분석한 선행 연구와 달리, Wang and Shepherd(2020)는 영국의 공공데이터 포털(data.gov.uk)에서 개방 중인 400개의 공공데이터에 대하여 완전성(Complete), 기본성(Primary), 적시성(Timely), 접근성(Accessible), 기계처리성(Machine Processable), 비차별성(non-Discriminatory), 비독점성(non-Proprietary), 무료 이용(Lincense-free) 등 8가지 원칙을 기반으로 만들어진 이용자 질의를 이용하여 공공데이터의 개방성을 측정하였다. Wang and Shepherd(2020)는 통계 정보와 같이 가공된 집합 정보(Aggregated

Information)는 원천 데이터(Raw Data)보다 활용성이 떨어지므로 공공데이터로서 부적절하다고 언급하고 있다. 이처럼, 국외 연구들은 자국의 공공데이터 포털을 중심으로 공공데이터 개방성을 측정하는 방법을 제안하고 있지만, 국내 연구들은 공공데이터 관련 국내 개방 정책의 실태 및 개선 방안을 제시하고 있다. Moon and Lee(2018)는 국내 공공데이터 관련 뉴스에 드러난 공공데이터 관련 이슈를 분석하여, 이슈별 공공기관과 정부 사업 현황을 분석하였다. Yun and Hyun(2019)은 공공데이터 포털을 통해 개방 중인 국가중점 데이터에 대하여 개방 데이터의 실태와 문제점을 분석하고 정책적 개선방안을 제시하고 있다. 공공데이터 개방 정책에 대한 문제점을 분석하기에는 국내 선행연구들이 적절하지만, 실제로 공공데이터를 이용하는 이용자들의 수요 분석을 통한 개방 정책의 방향성을 도출하는 데는 한계점이 존재한다.

지금까지 분석한 다양한 국내·외 공공데이터 선행연구를 분석해 보면, 연구마다 분석하고자 하는 연구 방법과 분석에 사용된 데이터, 도출된 결과가 본 연구와는 차이가 있다. 대부분의 연구가 공공데이터 포털 혹은 공공데이터 포털에서 개방 중인 공공데이터를 중심으로 각 나라의 데이터 개방성 혹은 개방 관련 정책을 분석하고 있다. 이를 통해, 진정한 의미에서 공공데이터로서 완전한 개방성을 갖추기에 부족한 점을 제시하고 있다. 가장 최근에 발표된 연구 중, 우선 본 연구와 가장 유사한 Cho and Ha(2020)는 공공데이터 포털에서 보유하고 있는 '공공데이터 제공신청' 및 '데이터 1번가' 정보에 대하여 구조적 토픽 모델링을 적용하여 국민·기업 등이 원하는 토픽을 도출하고, 각 분류별 제공신청 추이를 분석하였다. 공공데이터 수요가 높은 토픽에 대하여 중점적으로 개방해야 함을 제시하였으나, 본 연구의 분석 결과에 따르면 국민의 일상생활과 관련된 분야의 공공데이터는 이용자들의 공공데이터 수요 대비 초과하여 공공데이터가 개방되고 있다. 또한, 각 개인의 수요 측면에서 보면 이용자가 원하는 실제 공공데이터 개방은 즉각적으로 이뤄지지 않고 있다. 즉, 각 개인에게 필요한 공공데이터 수요를 파악하고, 그에 따른 공공데이터 개방 정책 수립(개방 우선순위 선정 등)은 아직 미흡한 실정이다. 본 연구에서는 키워드 네트워크 분석 기법을 적용하여, 공공데이터 수요 파악 및 이용 상황에 따른 공공데이터 수요의 변화 추이에 대하여 분석하고자 한다. 또한, 실제 운영 중인 공공데이터 포털의 대용량 텍스트 데이터를 활용하여, 좀 더 실증적인 분석 결과를 도출하고자 한다.

Ⅲ. 연구 설계

본 연구에서 키워드 네트워크 분석은 공공데이터 제공신청 수집에서부터 데이터 전처리, 추출된 키워드를 기반으로 키워드 네트워크를 구성하고, 키워드 네트워크 구조를 분석하는 과정을 포함한다. 이용자들의 관심이 높은 키워드에 대하여 발생 빈도 증가 추이를 파악하기 위해, 공공데이터 포털에서 수집된 공공데이터 제공신청으로부터 추출된 키워드 빈도 분석을 한다. 일반적으로 하나의 문서에 추출된 키워드의 발생 빈도가 높다고 하여, 해당 키워드들이 해당 문서의 의미를 잘 표현하지는 못한다. 키워드 동시 출현성 분석은 문서의 의미를 파악하기 힘들었던 키워드 빈도 분석의 문제점을 해결할 수 있다. 예를 들어, 하나의 문서에서 키워드 *mask*(마스크)의 발생 빈도수가 높다고 가정하자. 해당 문서가 *mask*(마스크)와 관련이 있다는 것은 명확하나, *mask sale*(마스크 판매)과 관련된 것인지 *mask price*(마스크 가격)와 관련된 것인지는 불분명하다. 키워드 동시 출현성 분석을 통해, *sale*(판매)이 동시 출현성이 높다면 해당 문서의 의미는 좀 더 명확해진다(즉, 복합어 *mask sale*(마스크 판매)과 관련된 문서로 판정). 본 연구는 문서에서 추출된 키워드 발생 빈도와 동시 출현성이 높은 키워드로 구성된 복합어를 문서의 주요 토픽으로 간주한다. 수집된 문서 집합 전체를 대상으로 추출된 토픽에 대하여 종속적 연관성을 분석을 수행하면, 공공데이터 이용자들의 잠재적 수요를 명확히 파악하는 것이 가능하다. 예를 들어, 토픽



〈그림 1〉 키워드 네트워크 분석 프레임워크
〈Fig. 1〉 Keyword network analysis framework

mask sale(마스크 판매)에 대하여 종속적 연관성 분석을 수행하면, *latitude & longitude of vendor*(제공업체 위·경도)에 대한 수요를 유추할 수 있다. 마지막으로, 토픽 연관성 분석을 통해 도출된 공공데이터 수요 정보가 현재 공공데이터 포털에 개방되어 있는지 비교를 위한 갭(Gap) 분석을 수행한다. 본 연구에 적용된 키워드 네트워크 분석 프레임워크를 도식화하면 〈그림 1〉과 같다. 본 연구는 키워드 네트워크 분석을 위한 도구로 KH Coder3(Higuchi, 2016)를 이용하였다.

1. 데이터 수집

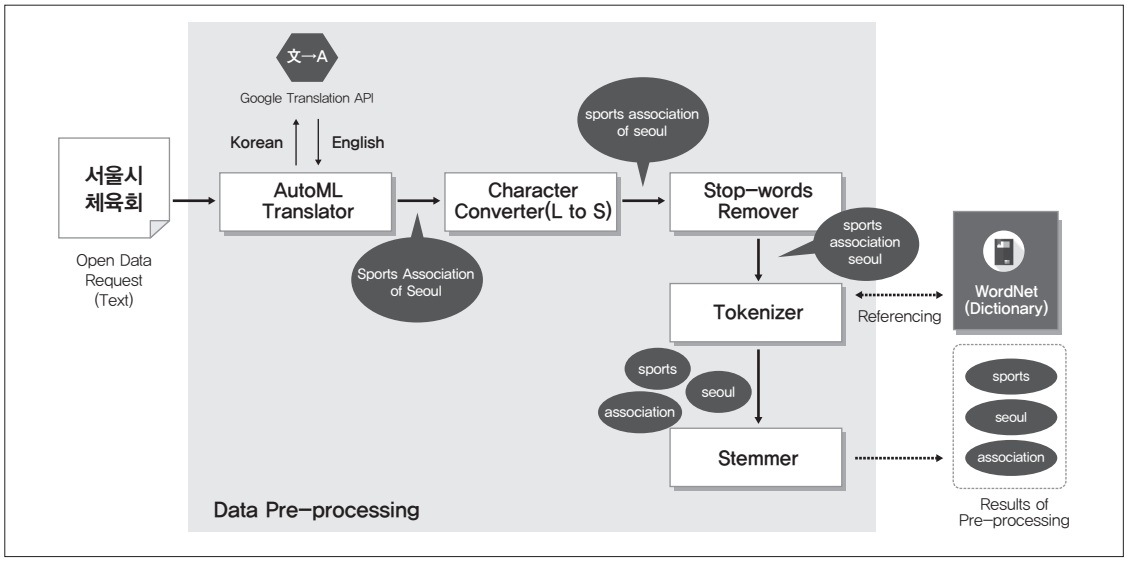
본 연구의 주요 목적은 상황별 공공데이터에 대한 수요 변화를 분석하고, 가장 적합한 공공데이터 정보를 예측하여 향후 공공데이터 개방 정책에 대한 방향성을 제시하는 것이다. 이를 위해, 본 연구에서는 공공데이터 포털에서 보유하고 있는 2019년 1월부터 2020년 8월까지 공공데이터 제공신청 정보 8,654건을 수집하였다. 2019년은 공공데이터 수요자 측면에서 특이 상황이 발생하지 않았으나, 2020년은 국내·외적으로 코로나19라는 특이 상황이 발생한 시기이다. 그러므로, 본 연구는 공공데이터 이용 상황(평상 vs. 긴급 상황)이 다른 2019년과 2020년에 공공데이터 수요 변화가 있을 것으로 가정한다. 또한, 공공데이터 제공신청 정보 중에는 비슷한 시기에 작성된 유사한 신청 정보들이 다수 존재한다. 본 연구는 공공데이터 제공신청자들의 개인정보는 수집 대상에서 제외하였으므로, 한 개인이 특정 토픽에 대하여 유사한 정보들에 대하여 제공신청한 중복 신청인지, 혹은 대중이 제공신청한 개별 신청인지 구분이 어렵다. 그러므로, 다수의 유사한 제공신청 정보에 대하여 중복 제거를 하지 않고, 해당 토픽에 대한 대중의 수요로 간주한다. 본 연구에서 수집·분석한 공공데이터 제공신청 정보는 실제 공공데이터 수요자가 공공데이터 포털을 통해서 신청한 정보이므로, 사례·설문·문헌 기반의 선행 실증 연구보다 더욱 정확한 수요 분석이 가능하다.

2. 데이터 전처리(Pre-processing)

공공데이터 포털을 통해 접수된 하나의 공공데이터 제공신청은 국문으로 작성된 하나의 문서(비정형 텍스트 데이터)와 같은 성격을 지닌다. 국문으로 작성된 문서에서 주요 키워드를 추출하기 위해서는 형태소 분석 과정을 거쳐야 한다. 하지만, 한글의 고유 특성(띄어쓰기 오류, 조사 등)으로 정확한 형태소 분석 결과를 얻기는 쉽지 않다. 예를 들어, '서울시체육회'와

같이 띄어쓰기 오류가 있는 복합 명사가 있을 때, 형태소 분석기에 따라 '서울시', '체육', '회(會)' 혹은 '서울', '시체', '육회'로 키워드가 추출될 수 있다. 그러므로, 본 연구에서는 형태소 분석에 대한 정확성과 텍스트 처리의 용이성을 확보하기 위해 국문을 영문으로 변환하여, 키워드 추출 및 수집을 수행한다. 구글(Google)에서 제공하는 번역 API(Google, 2020)를 활용하여 국문을 영문으로 변환한다. 위에서 언급한 예시인 '서울시체육회'에 대한 구글 번역 API를 적용하면, 결괏값 *Seoul, Sports, Association*이 추출된다. 영문은 국문과 달리 대문자와 소문자가 혼용되는 특성이 있다. 대·소문자의 혼용은 키워드 네트워크 분석 과정에서 키워드 발생 빈도 산정에 오류를 일으킬 수 있으므로, 본 연구는 모든 대문자를 소문자로 변환한다. 이후, 영문 문서에서 키워드를 추출하기 위해, 순차적으로 불용어(Stop Words) 제외 처리, 토큰화(Tokenization), 어간 추출(Stemming)을 수행한다 (Lee & Park, 2019). 불용어 제외 처리 과정은 전치사, 관사 등 너무 많이 등장하는 단어는 문서의 특징을 표현하는데 불필요한 단어들이므로 제외한다. 일반적으로 영문의 불용어로는 관사(a, an, the 등), 전치사(on, in, of 등), 특수 문자(느낌표, 구두점, 쉼표 등)가 존재한다. 토큰화는 문자열에서 단어를 분리하는 단계로 영문의 경우, 띄어쓰기를 중심으로 토큰화가 가능하다. 마지막으로 어간 추출은 단어의 기본 형태(예를 들어, 복수형 단어에서 단수형 단어 추출 등)를 추출하는 단계이다. 본 연구에서 수행하는 데이터 전처리 절차를 도식화하면 〈그림 2〉와 같다.

데이터 전처리 절차를 설명하기 위한 예시로, 입력 문서에 국문 '서울시체육회'가 포함되어 있다고 가정한다. 비록, 입력 문서에 띄어쓰기 오류가 있더라도, 구글 번역 API를 통해 띄어쓰기 오류가 수정되어 *Sports Association of Seoul*로 번역된다(실제로는 '*Seoul Sports Association*'으로 번역되나, 불용어 처리 과정 설명을 위해 본 예시는 전치사 'of'가 추가되어 *Sports Association of Seoul*로 번역된 것으로 가정한다). 번역된 영문에 대하여 대·소문자 변환 및 불용어(구두점, 쉼표 등 특수 문자 포함)를 제거한다. 본



〈그림 2〉 데이터 전처리 예시
〈Fig. 2〉 Example of data pre-processing

예시에서는 불용어인 전치사 'of'가 제거되어 *sports association seoul*이 된다. 본 연구에서는 불용어를 제외한 나머지 단일어들을 키워드(즉, 주요 단어)라 호칭한다. WordNet(Princeton, 2020) 참조를 기본으로 하되 일부 고유명사(기관명 등)를 추가하여, *sports association seoul*에 대하여 띄어쓰기를 기준으로 토큰화하면 키워드 *sports, association, seoul*이 추출된다. 어간 추출기(Stemmer)를 통해 각 키워드에 대하여 어간을 추출(불필요한 복수 표현 제거 등)하면, 최종적으로 *sport, association, seoul*과 같은 키워드 집합을 입력 문서("서울시체육회")의 전처리 결과로 얻게 된다. 일련의 데이터 전처리 과정을 마친 키워드 집합을 이용하여, 모든 문서(즉, 데이터 제공신청)에 대한 키워드 네트워크를 구성한다.

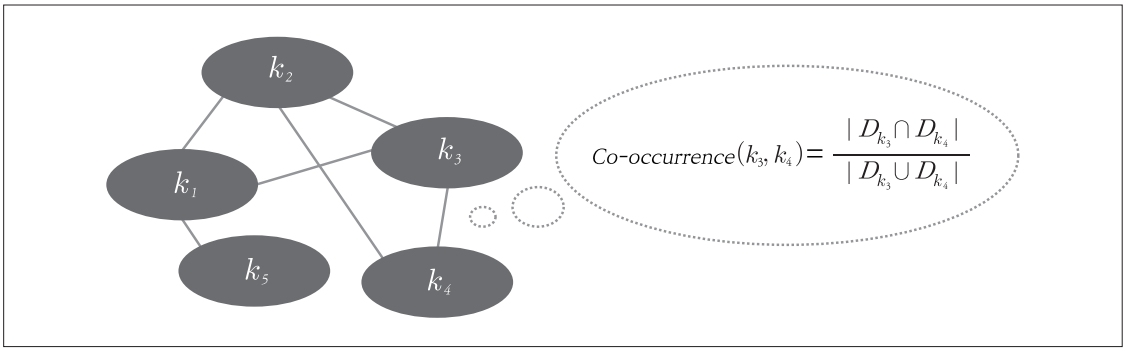
3. 키워드 네트워크 구성

앞서 수행한 데이터 전처리 과정을 거치게 되면, 하나의 문서(즉, 공공데이터 제공신청)는 여러 개의 키워드로 구성된 키워드 집합(Bag of Keywords)으로 표현된다. 전체 키워드 집합에서 동시에 출현하는 빈도가 높은 키워드 쌍은 상호 간에 높은 관련성을 지닌다. 따라서, 본 연구에서 구성하는 키워드 네트워크는 개별 문서를 구성하는 주요 키워드를 추출하여 결점(Node 혹은 Vertex)으로 표현하고, 동시에 출현하는 빈도가 높은 키워드 간의 관계를 연결선(Edge 혹은 Link)으로 표현한다. 〈그림 3〉은 하나의 문서에서 추출된 5개의 키워드를 결점으로 삼고, 동시 출현 빈도가 높은 키워드 쌍을 상호 연결하여 구성한 키워드 네트워크의 한 예시이다.

두 개의 키워드들 사이에 생성된 연결선은 방향성이 없으므로, 해당 키워드들은 상호 영향을 미치는 관계로 해석한다. 같은 방법을 수집된 문서 전체에 적용하여 하나의 키워드 네트워크를 만든다. 두 개의 키워드들의 동시 출현 빈도가 높을수록, 해당 키워드들은 높은 동시 출현성(co-Occurrence) 값을 갖는다. 본 연구에서 동시 출현성은 〈식 1〉과 같이 자카드 유사도(Jaccard Similarity)를 이용하여 측정한다(Baeza-Yates & Ribeiro-Neto, 1999).

$$Co\text{-}occurrence(k_i, k_j) = Jaccard\ Similarity(k_i, k_j) = \frac{|\ D_{k_i} \cap D_{k_j}\ |}{|\ D_{k_i} \cup D_{k_j}\ |},\ where\ \ k_i \neq k_j \qquad \langle 식\ 1 \rangle$$

〈식 1〉에서 D_{k_i}와 D_{k_j}는 각각 키워드 k_i와 k_j를 포함하고 있는 문서를 나타낸다. 그러므로, 〈식 1〉의 분모는 키워드 k_i또는 k_j를 포함하고 있는 전체 문서의 개수, 분자는 키워드 k_i와 k_j를 모두 포함하고 있는 문서의 개수



〈그림 3〉 키워드 네트워크 구성 예시
〈Fig. 3〉 Example of keyword network configuration

를 의미한다. 〈그림 3〉과 같이, 동시에 등장하는 모든 키워드 조합을 반영하여 네트워크를 구성하게 되면 결점 k_1, k_2, k_3 뿐만 아니라 결점 k_2, k_3, k_4 사이에 각각 네트워크 순환이 발생한다. 이러한 네트워크 순환은 네트워크 구조를 복잡하게 만들기 때문에, 분석 대상 네트워크 구조의 주축을 파악하기 어렵다. 그러므로, 본 연구에서는 네트워크에서 연결 가중치의 합이 최대가 되는 최대 신장 트리(maximum spanning tree)를 추출한다. 최대 신장 트리는 키워드 네트워크에서 동시 출현 빈도(즉, 연결선의 가중치)의 합이 최대가 되는 결점만을 연결한 하위 네트워크로서, 주어진 네트워크 구성을 간소화하는 장점이 있다.

Ⅳ. 데이터 분석

공공데이터 이용자들은 공공데이터 포털에서 미개방 중인 데이터에 대한 수요가 있을 때, 공공데이터 포털의 '공공데이터 제공신청' 서비스를 통해 각 공공기관에 공공데이터 제공을 요청할 수 있다. 하나의 공공데이터 제공신청 정보는 텍스트로 작성이 되어있으므로, 여러 개의 단어로 구성된 하나의 문서로 간주할 수 있다. 일반적으로 네트워크 구조(혹은 그래프 구조)는 하나 이상의 결점과 연결선으로 구성된다. 이때 하나의 연결선으로 맺어진 결점들은 상호 간에 영향을 미친다고 가정한다. 본 연구에서 수행하는 키워드 네트워크 분석에서는 키워드가 결점으로, 두 키워드의 간의 동시 출현성 및 종속적 연관성을 연결선으로 표현한다. 본 연구에서는 문서(즉, 공공데이터 제공신청)에서 발생 빈도가 높은 키워드를 추출하고, 이들 간의 관계를 분석에 활용하여 공공데이터 수요를 예측한다. 구체적으로, 하나의 문서에서 동시에 발생하는 키워드 즉, 동시 출현성 및 토픽 간의 종속적 연관성이 높은 키워드를 연결하여 전체적인 관계 구조를 살펴보기 위해, 키워드 간의 관계를 네트워크로 표현한다. 또한, 이용 상황에 따른 수요의 변화는 키워드 네트워크 구조의 변화를 통해 분석한다.

1. 키워드 빈도 분석

공공데이터 포털에서 수집된 공공데이터 제공신청 정보에 대하여 Ⅲ.2절에서 설명한 데이터 전처리 과정을 거쳐 5,528개의 키워드를 추출하였다. 이 중에서 분석 결과의 명확성을 위해 3,446개의 명사(Noun)만을 본 연구의 분석 대상으로 선정하였다. 가장 높은 빈도로 나타난 키워드는 *information*(정보)이며, *datum*(데이터; 단수형), *status*(현황), *number*(수), *city*(시) 등의 순이며, 상위 10개의 키워드와 발생 빈도는 정리하면 〈표 3〉과 같다. 〈표 3〉에서 높은 발생 빈도를 갖는 키워드들은 공공데이터 제공신청에 기술되는 관용적 표현(예를 들어, 제공신청 데이터명에 "OO 정보", "△△ 데이터", "XX 현황" 등)에서 추출되어 발생 빈도가 높다. 이들은 공공데이터 이용자들의 수요를 명확히 표현하는 키워드가 아니므로, 불용어로 간주하여 제거한다. 또한 분석 결과에 대하여 특정 지역

〈표 3〉 수집된 '공공데이터 제공신청'에서 발생 빈도가 높은 상위 10개 키워드
〈Table 3〉 Top-10 most frequent keywords in the collected OGD requests

Rank	Keyword	Frequency	Rank	Keyword	Frequency
1	information	1,735	6	seoul	470
2	datum	1,279	7	service	401
3	status	1,154	8	survey	380
4	number	754	9	area	252
5	city	534	10	korea	227

에 대한 편향된 효과(Biased Effect)를 제거하기 위해, *seoul*(서울) 등의 지명도 제거한다.

불용어가 모두 제거된 후, 연도별로 발생 빈도가 높은 키워드는 〈표 4〉와 같다.

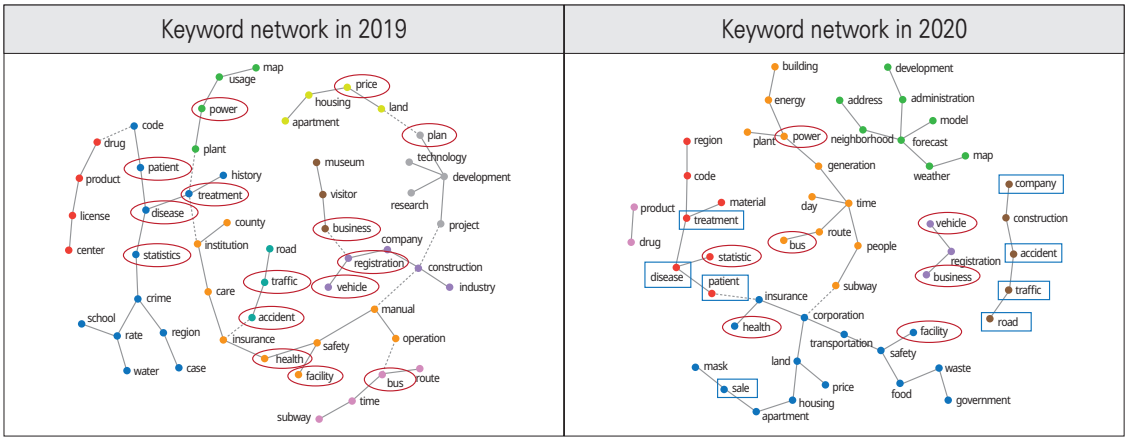
〈표 4〉에서 전년도 대비 발생 빈도 순위가 높아지거나, 신규로 상위 15위 안에 진입한 키워드는 이탤릭체로 표시하였다. 키워드 발생 빈도를 분석하면, 대부분의 공공데이터 이용자들이 필요로 하는 정보는 *vehicle*(차량), *business*(사업) 등과 관련된 해당 연도의 최신 *statistics*(통계) 정보임을 유추할 수 있다. 공공데이터 제공신청에서 2020년에 추출된 주요 키워드를 보면, *treatment*(치료), *company*(회사), *road*(도로), *sale*(판매) 등의 키워드를 포함하는 공공데이터 수요가 신규로 발생하였고, *disease*(질병), *accident*(사고), *traffic*(교통), *patient*(환자) 등의 키워드를 포함하는 공공데이터 수요가 전년 대비 증가(키워드 빈도 순위 3단계 이상 상승)한 것을 확인할 수 있다.

3. 키워드 동시 출현성 분석

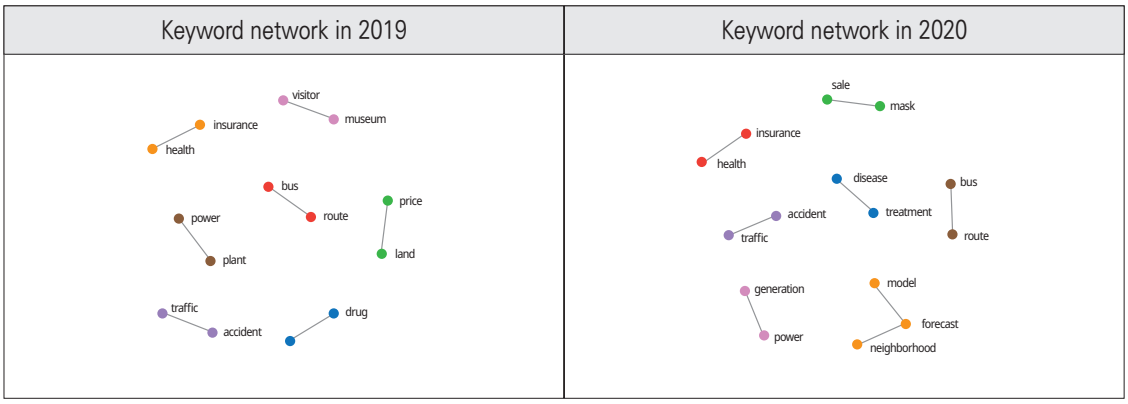
Palshikar(2007)에 따르면, 문서 집합에서 추출된 키워드 중에서 실제로 각 문서가 가진 고유성(혹은 토픽)을 잘 표현할 수 있는 핵심 키워드는 해당 문서에서 추출된 전체 키워드 대비 약 10%로 매우 적다. 그러므로, 문서가 가진 의미를 제대로 파악하기 위해서는 주어진 문서로부터 유의미한 키워드를 정확히 추출하는 것이 매우 중요하다. 본 절에서는 공공데이터 제공신청에서 발생하는 키워드 동시 출현성 분석을 통해 유의미한 키워드 간의 관계를 파악한다. 동시 출현 관계가 높은 키워드들은 언어학적으로 의미적 근접성(Semantic Proximity)이 높다는 것을 의미한다. 그러므로, 키워드 동시 출현성 분석은 문서의 의미를 파악하기 힘들었던 키워드 빈도 분석의 문제점을 해결할 수 있다. 만약, 공공데이터 제공신청에 대하여 키워드 분석 결과, 키워드 *sale*(판매)이 가장 출현 빈도가 높았다고 가정하자.

〈표 4〉 연도별 발생 빈도가 높은 상위 15개 키워드
〈Table 4〉 Top-15 most frequent keywords by year

2019			2020		
Rank	Keyword	Frequency	Rank	Keyword	Frequency
1	statistics	182	1	statistics	127
2	business	180	2	vehicle	121
3	vehicle	171	3	business	102
4	bus	167	4	power	87
5	facility	153	5	*disease*	74
6	power	153	6	*accident*	71
7	registration	117	7	*traffic*	71
8	visitor	114	8	*patient*	70
9	disease	113	9	bus	69
10	accident	112	10	facility	67
11	patient	111	11	*treatment*	67
12	traffic	110	12	*company*	66
13	health	104	13	*road*	65
14	plan	95	14	*sale*	62
15	price	95	15	health	61



(a) 연도별 키워드 동시 출현성 네트워크
(a) keyword co-occurrence network by year



(b) 높은 동시 출현성 값을 갖는 상위 7개 키워드 네트워크
(b) Top-7 keyword networks with high co-occurrence

〈그림 4〉 키워드 동시 출현성 분석 결과
〈Fig. 4〉 A result of keyword co-occurrence analysis

이 경우, 키워드 *sale*(판매)만으로는 무슨 공공데이터를 이용자가 원하는지 정확히 파악하기 쉽지 않다. 하지만, 〈그림 4〉와 같이 키워드 동시 출현성 분석을 하면, 키워드 *sale*(판매)은 키워드 *mask*(마스크)와 동시 출현성이 매우 높음을 알 수 있다(〈그림 4(b)〉의 키워드 동시 출현성 분석 결과 참조). 이를 통해, 문서의 의미를 나타내는 복합어 *mask sale*(마스크 판매)을 도출할 수 있다. 일반적으로, 복합어는 단일어로 구성된 키워드 대비 좀 더 구체적인 문서의 의미를 표출하는 것이 가능하므로, 본 연구는 Ⅲ장에서 언급한 것과 같이 복합어를 문서의 주요 토픽으로 간주한다.

〈그림 4(a)〉는 〈식 1〉에 의해 도출된 키워드 동시 출현성 값을 기준으로 연도별 키워드 네트워크를 도식화하고, 〈그림 4(b)〉는 키워드 네트워크의 토픽 변화 추

이를 분석하기 위해 키워드 동시 출현성이 높은 상위 7개의 키워드 네트워크를 도식화한 것이다. 독자들의 가독성을 높이기 위해, 〈표 4〉의 발생 빈도가 높은 키워드들은 붉은 원으로 표시하였으며, 2019년 대비 2020년에 발생 빈도 순위가 높아진 키워드들은 파란 네모로 표시하였다. 〈그림 4(b)〉를 보면, 2019년의 경우, *health insurance*(건강보험), *museum visitor*(박물관 방문객), *traffic accident*(교통사고), *bus route*(버스노선), *land price*(토지 가격), *power plant*(발전소), *drug product*(약품) 등 일상생활과 밀접한 관련성을 갖는 키워드들이 높은 동시 출현성을 갖는다. 반면, 2020년의 경우, *health insurance*(건강보험), *traffic accident*(교통사고), *bus route*(버스노선), *power generation*(발전), *neighborhood forecast*(동네예보) 등 일상생활과 밀접한 관련성을 갖는 키워드뿐만 아니라, 사회적 이슈(코로나19 등)와 밀접한 연관성을 갖는 키워드인 *mask sale*(마스크 판매), *disease treatment*(질병 치료) 등이 전년 대비 높은 키워드 동시 출현성을 갖는다.

본 절에서 수행한 키워드 동시 출현성 분석 결과를 보면, 공공데이터 제공신청은 공공데이터 이용자들이 겪는 사회적 이슈(혹은 이용자 상황)와 밀접한 연관을 가지고 변화한다. 그러므로, 이용자 수요가 높은 데이터를 신속하게 제공하기 위해서는 사회적 이슈 및 이용자 관심 사항들에 대한 지속적 모니터링 및 파악이 필요하다.

4. 토픽 연관 분석

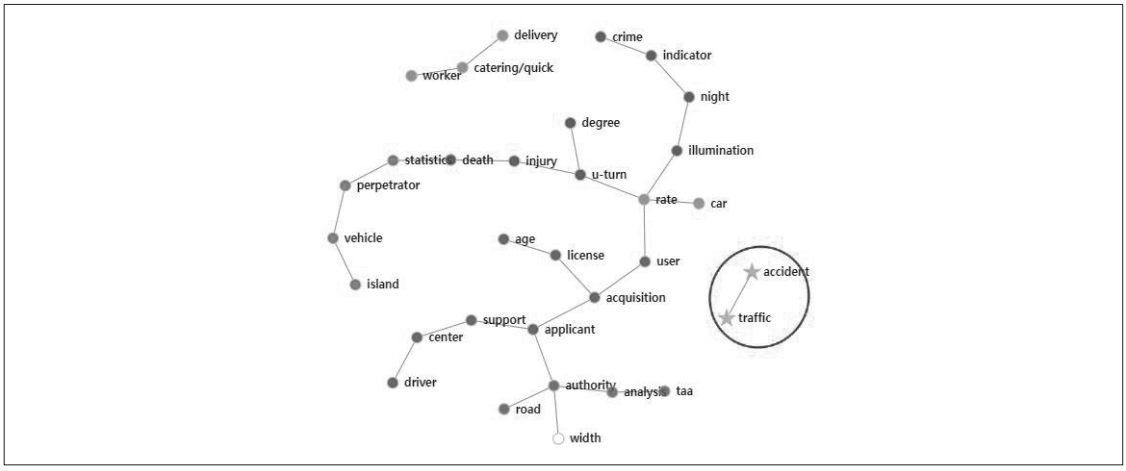
본 절에서는 키워드 동시 출현성 분석을 통해 도출된 특정 토픽(즉, 동시 출현성이 높은 키워드 쌍으로 구성된 복합어)에 대하여 연관 분석을 수행한다. 연관 분석은 통계학적 분석 방법으로 조건부 확률을 이용하며, 특정 조건에 대하여 종속된 연관 정보를 예측하는 것이 가능하다. 본 연구에서 두 키워드 간의 연관성은 〈식 2〉와 같이 조건부 확률을 이용하여 측정한다.

$$Association(T_i, T_j) = Pr(T_i \mid T_j) = \frac{Pr(T_i \cap T_j)}{Pr(T_j)} \quad \langle 식\ 2 \rangle$$

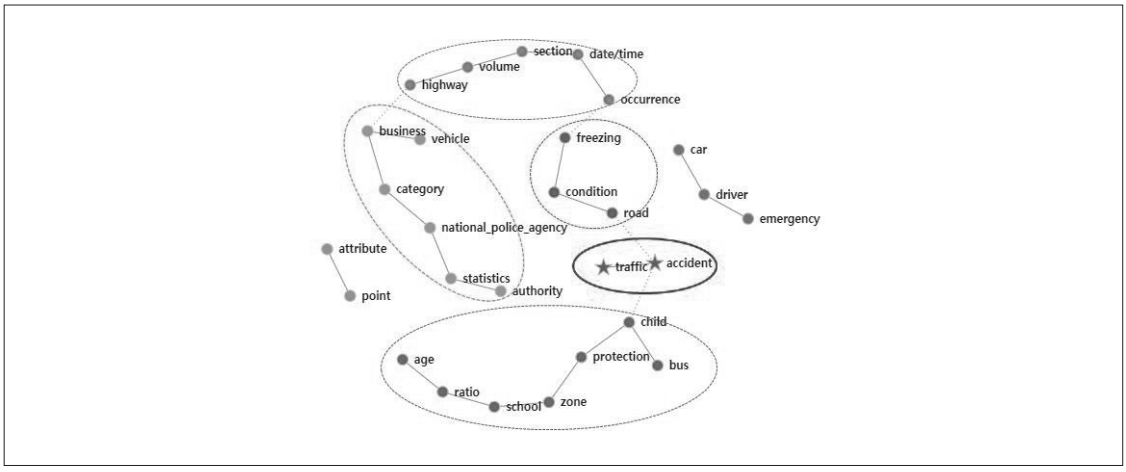
〈식 2〉에서 조건부 확률 $Pr(T_i \mid T_i)$는 토픽 T_j가 조건으로 주어졌을 때, 토픽 T_i가 발생할 확률을 의미한다. 〈식 2〉의 분모는 토픽 T_j와 관련된 문서의 발생 확률, 분자는 토픽 T_i, T_j와 모두 관련된 문서의 발생 확률을 의미한다. 그러므로, 〈식 2〉를 통해, 토픽 T_j에 대한 토픽 T_i의 종속적 연관성을 측정하는 것이 가능하다. 이를 본 연구에 적용하면, 특정 토픽이 조건으로 주어졌을 때, 그에 대한 종속적 연관 정보(공공데이터에 대한 추가 수요 정보)를 예측할 수 있다. 본 절에서는, 2019년과 2020년에 공통으로 높은 동시 출현성 값을 갖는 토픽 *traffic accident*(교통사고)와 2020년에 신규로 높은 동시 출현성 값을 갖는 토픽 *mask sale*(마스크 판매)을 중심으로 설명한다. 앞서 수행한 키워드 빈도 분석과 동시 출현성 빈도를 통해 공공데이터 이용자들은 토픽 *traffic accident*(교통사고)와 토픽 *mask sale*(마스크 판매)에 대한 공공데이터 수요가 높은 것을 알 수 있다. 하지만, 구체적으로 해당 토픽들과 관련하여 어떤 공공데이터를 원하는지 구체적으로 파악하기는 쉽지 않다. 특히, *mask sale*(마스크 판매)은 최근 코로나19 이슈와 맞물려서 공공데이터 제공신청이 급증하여 높은 빈도로 출현하는 토픽이다.

우선, 토픽 *traffic accident*(교통사고)에 대하여 연도별 키워드 분석 결과를 도식화하면 〈그림 5〉와 같다. 〈그림 5〉에서 조건 토픽은 붉은 원안에 별표로 표시하고, 토픽 연관 분석 관계와 키워드 동시 출현성 분석 관계는 각각 점선, 실선으로 나타낸다. 2019년의 경우, *traffic accident*(교통사고)와 동시 출현 빈도가 높은 키워드 *road*(도로), *perpetrator*(가해자), *driver*(운전자), *age*(연령), *injury*(상해), *license*(면허)가 추출되었으며, 이들 간의 실선으로 연결된 동시 출현성 관계를 확인할 수 있다. 하지만, 토픽 *traffic accident*(교통사고)와 다른 토픽 간의 점선으로 연결된 종속적 연관성은 존재하지 않는다. *traffic accident perpetra-*

tor vehicle(교통사고 가해자 차량), *traffic accident license acquistion*(교통사고 면허 취득) 등과 같이 교통사고 정보에 대한 공공데이터 수요는 높으나, 주어진 토픽 *traffic accident*(교통사고)와 종속적 연관 관계를 갖는 공공데이터 제공신청 정보는 실제로 찾기 어렵다. 반면, 2020년의 경우, 4개의 토픽(점선 원)이 토픽 *traffic accident*(교통사고)와 직·간접적으로 종속적 연관 관계를 맺고 있다. 분석 결과를 보면, *traffic accident related to road freezing condition*(도로 결빙상태와 관련된 교통사고), *section and date/time of occurrence related to traffic accident*(교통사고와 관련된 사고 발생 구간 및 일시), *highway traffic accident related to business vehicle*(영업용 차량과 관련된 고속도로 교통사고), *traffic accident related to child protection(or school) zone*(어린이보호구역과 관련된 교통사고) 등이며, 이들은 토픽



(a) Topic association analysis in 2019



(b) Topic association analysis in 2020

〈그림 5〉 연도별 토픽('교통사고') 연관 분석 예시
〈Fig. 5〉 Example of topic ('traffic accident') association analysis by year

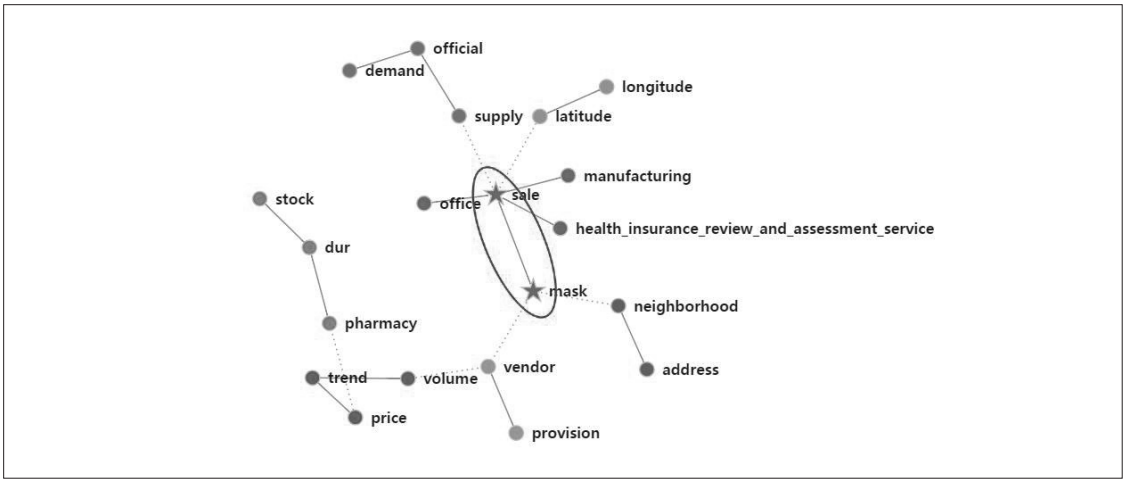
traffic accident(교통사고)와 직·간접적으로 종속적 연관 관계를 내포하고 있다. 즉, 토픽 연관 분석 결과에서 도출된 정보에 대하여 종속적 연관 관계를 분석함으로써, 공공데이터 이용자들의 수요 정보를 정확히 예측하는 것이 가능하다.

다음으로, 토픽 *mask sale*(마스크 판매)에 대하여 키워드 연관 분석을 수행한다. 〈그림 6〉은 *mask sale*(마스크 판매)을 조건으로 하였을 때, 토픽 연관 분석을 수행한 예시이다.

분석 결과를 보면, 토픽 *mask sale*(마스크 판매)은 *latitude*(위도), *longitude*(경도), *price*(가격), *stock*(보유) 및 *vendor*(공급업체) 등의 키워드와 밀접한 연관 관계를 갖는다. 즉, 이용자들은 '마스크 판매'와 토픽 연관성이 높은 '마스크 판매처 위치', '마스크 판매 가격', '마스크 판매 가능한 보유량', '마스크 판매량' 등에 대하여 공공데이터 수요가 높은 것을 예측할 수 있다. 사회적 파급력이 높은 공공데이터 수요 정보의 경우, 공공데이터 이용자들이 원하는 정보를 정확히 파악하여, 신속하게 적시에 제공할 필요가 있다. 실례로, 공적 마스크 정보를 적시에 개방하여 민·관 협업으로 코로나19에 대응한 K-방역의 성공 사례와 같이, 국민의 관심이 높은 사회적 이슈에 대하여 이용자들이 원하는 공공데이터를 정확히 파악하여 적시에 제공하는 것은 긴급 상황 발생 시 위기 상황 극복을 위해 매우 중요하다. 본 연구에서 제안하는 키워드 네트워크 분석 프레임워크를 활용하여 공공데이터 수요 예측을 정확하게 수행함으로써, 적시성 있는 공공데이터 개방을 추진할 수 있을 것으로 기대한다.

5. 공공데이터 제공신청 기반 공공데이터 수요 예측 및 개방 현황 분석

본 절에서는 앞서 수행한 키워드 동시 출현성 분석과 토픽 연관 분석 결과를 기반으로 현재 공공데이터 포털에서 개방 중인 공공데이터와의 갭 분석을 수행한다. 2019년부터 2020년까지 공공데이터 이용자들이 신청한 공공데이터 정보와 현재 공공데이터 포털에서 개방 중인 공공데이터 정보를 분석 대상으로 선정하였다. 본 연구에서 수행한 키워드 빈도 분석, 키워드 동시 출현성 분석, 토픽 연관 분석, 개방과 수요 갭 분석을 반복적으로 수행함으로써, 민간의 수요는 높으

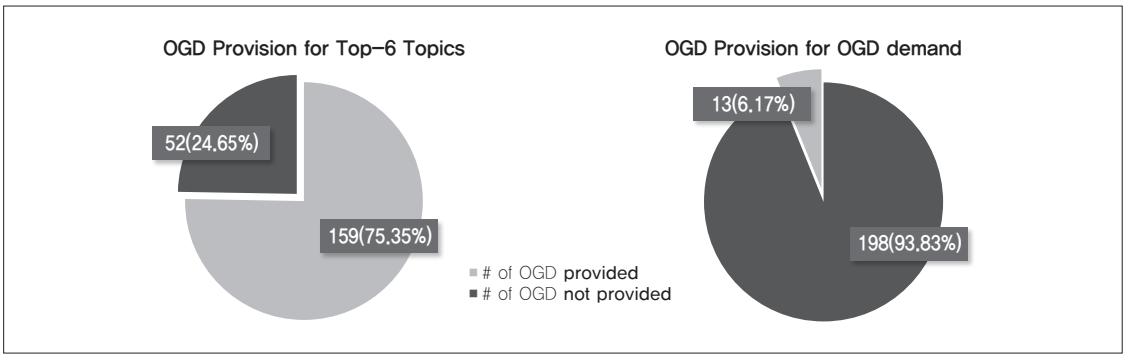


〈그림 6〉 토픽 연관 분석 예시('마스크 판매')
〈Fig. 6〉 Example of topic association analysis ('mask sale')

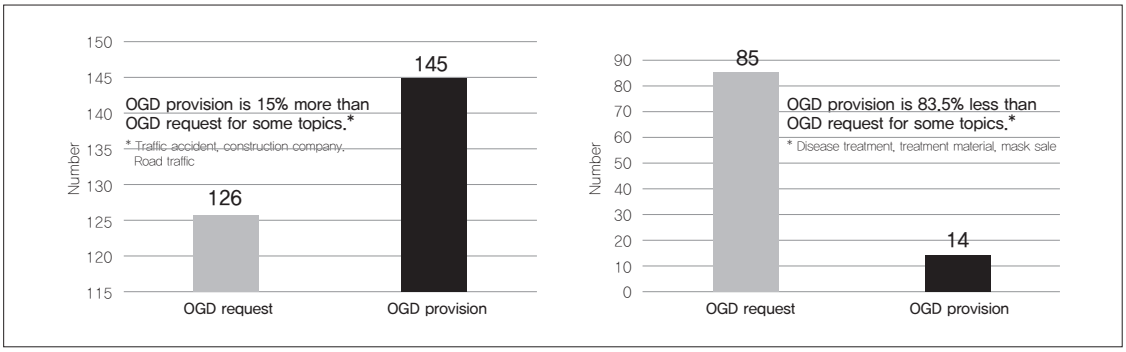
나 아직 미개방 중인 공공데이터 정보를 파악하고, 공공데이터 개방에 대한 우선순위 선정에 이바지할 수 있을 것으로 기대한다. 〈표 5〉는 〈표 4〉에서 발생 빈도가 급증한 7개의 키워드에 대하여 동시 출현성 분석 및 연관 분석을 수행한 결과이다. Ⅳ.2절 및 Ⅳ.3절에서 설명한 키워드 동시 출현성 분석 및 토픽 연관 분석을 통해, 공공데이터 제공신청에서 가장 수요가 높은 공공데이터 정보를 추출하였다. 또한, 토픽 연관 분석을 통해 도출된 공공데이터 이용자들의 수요에 대한 개방 현황을 정리하면 〈표 5〉와 같다. 〈표 5〉에 의하면, 키워드 동시 출현성 분석을 통해 도출된 6개의 토픽(*disease treatment, traffic accident, treatment material, construction company, road traffic, mask sale*)과 관련하여 211건의 공공데이터 제공신청이 있었으며, 해당 토픽과 관련된 공공데이터 159개가 공공데이터 포털에 개방되어 약 75.35%의 개방률을 나타내고 있다. 또한, 토픽 연관 분석을 통해 이용자들이 실제로 원하는 것으로 예측된 공공데이터는 겨우 13개가 공공데이터 포털에 개방되고 있어 공공데이터 개방률은 6.17%로 저조하다. 〈표 5〉의 분석 결과를 기반으로, 공공데이터 토픽(즉, 키워드 동시 출현성 분석 결과)과 관련된 공공데이터 개방 현황과 공공데이터 수요(즉, 토픽 연관 분석 결과)와 관련된 공공데이터 개방 현황을 각각 요약하여 도식화하면 〈그림 7〉과 같다. 토픽 *traffic accident*(교통사고), *construction company*(건설업체), *road traffic*(도로교통)과 같이 국민의 일상생활과 밀접한 관련성을 갖는 경우, 제공신청 건수 대비 개방 건수가 약 15% 많다. 실제로 토픽 *traffic accident*(교통사고)와 *road traffic*(도로교통)의 경우,

〈표 5〉 공공데이터 토픽과 수요에 대한 공공데이터 개방 현황
〈Table 5〉 OGD provision status for OGD topics and OGD demand

Single Keyword	Keyword Co-occurrence Analysis			Topic Association Analysis	
	OGD Topic	# of OGD requested	# of OGD provided	OGD Demand	# of OGD provided
disease	disease treatment	34	2	N/A	-
accident	traffic accident	79	91	child-protection-zone-school	10
traffic				road-condition-freezing	1
treatment	treatment material	35	9	vacuum-pressure-claim	1
				valve-disease	0
				billing-statistics	1
company	construction company	17	15	power-plant	0
				order	0
road	road traffic	30	39	safety-paper	0
sale	mask sale	16	3	vendor-provision	0
				neighborhood-address	0
				latitude-longitude	0
Total number of OGD		211	159		13
Percentage of OGD provision		75.35%			6.17%



〈그림 7〉 공공데이터 토픽과 수요에 대한 공공데이터 개방 현황
〈Fig. 7〉 OGD provision status for OGD topics and OGD demand



〈그림 8〉 토픽별 공공데이터 제공신청과 개방 현황 비교
〈Fig. 8〉 Comparision of OGD request & provision status by topics

이용자들의 제공신청 건수보다 공공데이터 포털을 통해 개방 중인 건수가 더 높은 것을 확인할 수 있다.

반면에, 토픽 *disease treatment*(질병치료), *treatment material*(치료물질), *mask sale*(마스크 판매)와 같이 사회적 이슈(2020년의 코로나19 등)와 관련된 데이터는 제공신청 대비 개방 건수가 약 83.5% 적다. 〈표 5〉에서 공공데이터 이용 상황과 관련된 토픽별 공공데이터 제공신청 건수 대비 개방 건수를 비교하여 도식화하면 〈그림 8〉과 같다. 이처럼, 이용자들이 원하는 공공데이터 개방률이 낮은 원인은 공공데이터를 개방할 때 수요자 관점에서 원하는 공공데이터를 상세하게 파악하기보다는 이용자 수요가 높은 토픽과 관련된 공공데이터 중에서 공공기관 담당자들이 개방하기 쉬운 공공데이터를 우선 개방하기 때문이다.

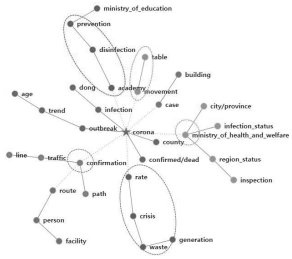
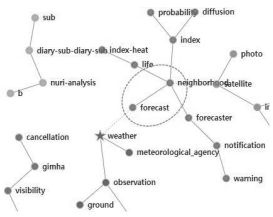
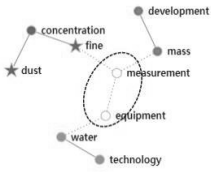
〈그림 9〉는 〈표 5〉에서 공공데이터 수요보다 공공데이터 개방 건수가 높은 토픽 *traffic accident*(교통사고)에 대하여, 토픽 연관 분석 결과인 '*child-protection-zone-school*'을 공공데이터 포털에서 실제 검색한 결과의 예시이다. 토픽 연관 분석을 통해 도출된 공공데이터 수요 정보에 대하여 공공데이터 포털에서 현재 개방 중인 공공데이터를 찾기 위해, 도출된 토픽 키워드를 활용하여 AND, OR로 검색식을 구성하여 공공데이터 포털에서 검색을 수행하였다.

〈그림 9〉 공공데이터 포털(data.go.kr) 내 개방 데이터 검색 예시
〈Fig. 9〉 Example of OGD search in OGD portal (data.go.kr)

6. 인기 검색어 기반 공공데이터 수요 예측 및 개방 현황 분석

앞서 설명한 공공데이터 제공신청뿐만 아니라, 공공데이터 포털에서 제공 중인 인기 검색어 정보를 통해서도 실시간으로 공공데이터 수요를 분석하는 것이 가능하다. 공공데이터 포털은 이용자들이 공공데이터 포털에서 수행한 검색 로그를 기반으로 매일 인기 검색어 상위 5개에 대한 정보를 제공한다. 공공데이터 포털에서 제공 중인 인기 검색어 목록(2020.8.25. 기준) 및 키워드 연관 분석을 통해 도출된 공공데이터 수요 예측을 하면 〈표 6〉와 같다. 일반적으로, 인기 검색어는 공공데이터에 대한 실시간 검색 수요이므로, 공공데이터 제공신청을 통한 개방 수요와는 차이가 있다. 〈표 6〉에서 보이는 것과 같이 인기 검색어는 1~2개의 키워드로 구성되어 있으므로, 공공데이터 제공신청보다 이용자의 수요를 정확하게 파악하기 쉽지 않다는 단점이 있다. 하지만, 공공데이터 이용자들의 실시간 수요를 적시성 있게 파악하기 쉽다는 장점이 있다. 〈표 6〉에서 연관 분석을 통해 도출된 공공데이터 수요는 점선 원으로 색상을 달리 표시하였으며, 색상별로 "OGD Demand(공공데이터 수요)" 항목에 표시하였고, 연관 분석 결과 분기 노드가 있으면, 해당 노드의 키워드만 명시하였다. 예를 들어, 인기 검색어 *corona*(코로나)의 경우, *confirmation*(확진)에서 트리 분기(녹색 원)가 발생한다. 이 경우 모든 분기 경로를 명시하지 않고, 분기 노드인 *confirmation*(확진)만 〈표 6〉의 OGD Demand 항목에 명시하였다. 분석 결과를 보면, 인기 검색어 *corona*(코로나)에 대하여 추가로 *academy-disinfection-prevention*(학원별 코로나 방역 및 소독), *movement-table*(동선표), *conrfirmation*(확진자 관련 정보), *rate-crisis-waste-generaton*(위기 상황 쓰레기 배출 비율), *mistry_of_health _and_welfare*(보건복지부 관련 정보) 등에 대한 공공데이터 수요를 예측할 수 있다. 또한, 인기 검색어 *weather*(날씨)와 관련하여 토픽 *forecast-neighborhood*(동네예보), 인기 검색어 *fine dust*(미세먼지) 관련하여 토픽 *measurement-equipment*(측정 장비)에 대한 수요를 각각 도출하였다. 인기 검색

〈표 6〉 인기 검색어에 대한 공공데이터 개방 현황
〈Table 6〉 OGD provision status for popular search keywords

Rank	Search Keyword	Association Analysis	OGD Demand	# of OGD provided
1	corona		*academy-disinfection-prevention*	0
			movement-table	0
			confirmation	13
			rate-crisis-waste-generation	0
			ministry_of_health_and_welfare	8
2	weather		*forecast-neighborhood*	9
3	fine dust		*measurement-equipment*	0
4	subway	N/A	N/A	N/A
5	defense		*restriction-disposition-acquisition-qualification*	2
			disclousure-standard	2
			project-pmbok	0
			department-px-product	2
			cooperation	0
Total number of OGD provided by the Korean OGD portal				28

어 *subway*(지하철)는 토픽 연관 분석을 통해 종속적 연관 관계가 도출되지 않아, 개방 현황 분석에서 제외하였다. 마지막으로, 인기 검색어 *defense*(국방)와 관련해서는 종속 관계를 확장해 보면 토픽 *restriction-disposition-acquisition-qualification*(자격 취득 처분 제한), *disclouser -standard*(공개 표준), *department-px-product*(PX 상품 부서) 등에 대한 정보 수요가 예측되었다. 앞서 언급한 것처럼, 이들은 검색 수요에 대하여 토픽 연관 분석을 수행한 것이므로, 공공데이터 제공신청을 통한 개방 수요와 달리 공공데

이터 포털을 통해 이미 개방 중인 경우가 많다. 실제로, 공공데이터 제공신청에 대한 공공데이터 포털에 개방 중인 공공데이터 개방 건수는 〈표 5〉에서와 같이 13건이지만, 인기 검색어에 대한 공공데이터 개방 건수는 〈표 6〉에서와 같이 28건이다. 〈표 5〉의 경우 실제 공공데이터 제공신청 건수 대비 수요가 예측된 공공데이터의 개방 건수를 백분율로 표현하는 것이 가능하지만, 〈표 6〉의 경우 얼마나 많은 이용자가 각 인기 검색어를 통해 공공데이터 요청하였는지를 파악하기 쉽지 않으므로, 제공신청 건수 대비 개방 건수를 백분율로 표현하는 것이 불가능하다. 공공데이터 제공신청 대비 인기 검색어에 대한 개방 건수는 높은 편이지만, 수요자 관점에서 보면 여전히 미개방 중인 공공데이터가 많이 존재한다.

Ⅴ. 결론

본 연구는 공공데이터 제공신청 데이터에 대하여 키워드 빈도 분석, 키워드 동시 출현성 분석, 키워드 연관 분석을 순차적으로 적용하여 키워드 네트워크를 구성하여, 공공데이터 이용자들의 수요를 예측하였다. 본 연구의 분석 결과 및 시사점을 요약하면 다음과 같다.

첫째, 공공데이터 포털을 통해 접수된 공공데이터 제공신청 내용에 대하여 가장 키워드 빈도가 높은 키워드는 *staticstic*(통계), *vehicle*(차량), *business*(사업) 등에 대한 공공데이터 수요가 높다. 2019년 대비 2020년에는 사회적 이슈(코로나19 등)와 관련된 키워드 *disease*(질병), *patient*(환자), *treatment*(치료)에 대한 출현 빈도가 급증한 것도 확인할 수 있다. 그러므로, 사회적 이슈와 관련된 키워드를 미리 발굴·분석하여, 민간의 공공데이터 수요에 대하여 선제적으로 대응할 필요가 있다.

둘째, 키워드 동시 출현성 분석을 통해, 키워드 빈도 분석을 통해 도출된 키워드에 대한 의미를 구체화할 수 있다. 이는 동시 출현 관계가 높을수록, 두 키워드 간의 의미적 근접성이 높다는 것은 의미하기 때문이다. 실제로, 키워드 출현 빈도가 높으나 의미적 모호성을 갖는 키워드 *accident*(사고), *sale*(판매)에 대하여, 키워드 동시 출현성 분석을 통해, 의미적 모호성이 제거된 토픽 *traffic accident*(교통사고), *mask sale*(마스크 판매)을 확인할 수 있다.

마지막으로, 키워드 동시 출현성 분석을 통해 도출된 토픽에 대하여, 토픽 연관 분석을 수행함으로써 종속적 연관 관계가 있는 토픽을 유추하였다. 유추된 토픽들을 활용하여, 해당 토픽과 관련하여 공공데이터 이용자들의 수요에 대한 정확한 예측이 가능하다. 앞서 언급한 것처럼, 사회적 관심이 높은 공공데이터 정보의 경우, 공공데이터 이용자들이 원하는 정보를 정확히 파악하여, 적시성 있게 제공하는 것은 매우 중요하다. 추가로, 공공데이터 이용자들의 수요에 적시성 있고 기민하게 대응하기 위해, 공공데이터 제공신청뿐만 아니라 공공데이터 인기 검색어에 대하여 토픽 연관 분석을 수행하였다. 일반적으로, 공공데이터 제공신청은 현재 공공데이터 포털에서 미개방 중인 데이터에 대한 개방 수요인 반면, 인기 검색어는 공공데이터 포털에서 개방(혹은 미개방) 중인 데이터에 대한 검색 수요이다. 검색 수요를 통해 공공데이터 이용자들의 관심 사항 혹은 수요 추이를 실시간으로 파악하는 것이 가능하다. 그러므로, 개방 수요뿐만 아니라 검색 수요를 함께 분석함으로써, 미개방 중인 공공데이터에 대한 중장기적 공공데이터 개방 전략뿐만 아니라 실시간 공공데이터 개방 전략 수립도 가능할 것으로 기대한다. 특히, 마스크 정보를 제공함으로써 코로나19에 대응하는 K-방역이 성공한 사례를 보면, 공공데이터 이용자들이 원하는 정보를 정확히 파악하고, 관련된 공공데이터를 신속하고, 기민하게 개방하는 것은 매우 중요하다.

본 연구는 공공데이터 포털이 보유하고 있는 대규모의 공공데이터 제공신청 정보에 대하여 키워드 네트워크 분석 기법을 순차적으로 적용하여 공공데이터 이용 상황(평상, 긴급 상황)별 정확한 수요를 예측하였다. 실제로 대국민 서비스를 제공 중인 공공데이터 포털에서 보유하고 있는 대용량 텍스트 데이터를 대상으로, 선

행 연구와 달리 공공데이터 수요에 대한 토픽뿐만 아니라 구체적으로 키워드 레벨에서 이용 상황별 공공데이터 수요 분석을 수행한 첫 연구라는 점에서 큰 의의가 있다. 아직도 많은 공공기관(중앙부처·지방자치단체·산하기관)이 수요자 중심의 공공데이터 개방 정책 수립을 위한 참고자료 확보를 위해 공공데이터 수요 조사(온·오프라인 설문조사, 기업간담회 등)를 수행하고 있다. 이러한 조사는 제한된 인원의 참여와 시·공간의 제약을 수반하고, 수요조사 참여자들이 공공데이터에 대한 실제 수요가 존재하는지가 불명확하므로 비용적인 측면 및 정확도 측면에서도 문제점이 많다. 반면, 본 연구는 공공데이터 수요가 있는 실제 이용자들이 제공신청한 비정형 텍스트 데이터를 대상으로 하였기 때문에 비용적인 측면 및 정확도 측면에서도 유효한 공공데이터 수요 정보를 예측하였다. 그러므로, 본 연구에서 제안한 키워드 네트워크 기반 분석 프레임워크는 공공데이터 수요를 파악하기 위해 실무적으로 적용이 가능한 분석 방법이므로, 향후 정부의 공공데이터 개방 정책 수립에 널리 활용될 수 있을 것으로 기대된다. 다만, 본 연구의 결과는 이용 상황에 따른 공공데이터 수요 변화를 분석하기 위해 2019년부터 2020년까지의 공공데이터 제공신청 정보를 기반으로 키워드 네트워크를 구성하였고, 이로부터 주요 토픽 및 키워드에 대한 공공데이터 수요를 예측하였으나, 좀 더 정확한 키워드 네트워크 구성 및 수요 분석·예측을 위해서는 더 많은 기간의 공공데이터 제공신청 정보를 확보하여 분석할 필요가 있다. 수집된 데이터양이 많아질수록 편향된 분석 결과가 나올 가능성은 줄어들기 때문이다. 또한, 현재 공공데이터 포털을 통해서 국민과 기업이 데이터를 요구하는 방법은 본 연구에서 분석 대상으로 선택한 공공데이터 제공신청 정보도 있지만, '데이터 1번가' 정보도 존재한다. 공공데이터 제공신청을 위해서는 공공데이터 제공신청서를 작성하고, 공공데이터 포털에 회원 가입 및 로그인을 해야 한다. 이러한 작은 불편 사항을 개선하기 위해서 만들어진 서비스가 데이터 1번가이다. 데이터 1번가는 공공데이터법과 같은 법적 근거에 따라 운영하는 제도는 아니지만, 국민이 필요한 데이터를 실시간으로 개방 요청하고 해당 기관이 온라인으로 즉시 답변해 주는 방식으로 2018년부터 운영되고 있다. 따라서, 다양한 채널(공공데이터 제공신청, 데이터 1번가 등)을 통해 접수되는 민간 이용자들의 공공데이터 요구사항을 분석하여, 본 연구 결과와 비교할 필요가 있다. 또한, 온라인 포털(네이버, 다음 등), 뉴스 등을 통해 수집할 수 있는 인기 검색어를 기반으로 사회적 이슈에 관한 변화를 모니터링하는 것이 가능하다. 그러므로, 사회적 이슈와 관련성이 높은 키워드를 중심으로 공공데이터 수요 예측과 개방 현황에 대한 비교 분석을 향후 연구 과제로 수행할 필요가 있다.

■ References

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York: ACM Press.

Blei, D. & Lafferty. J. (2007). "A Correlated Topic Model of Science." *The Annuals of Applied Statistics, 1*(1), 17-35

Chatfield, A. & Reddick C. (2017), "A Longitudinal Cross-sector Analysis of Open Data Portal Service Capability: the Case of Australian Local Governments." *Government Information Quarterly, 34*(2), 231-243.

Cho, S. & Ha, S. (2020). "Analysis of Open Government Data Demand Using Structural Topic Modeling." *Journal of Information Technology and Architecture, 17*(2), 103-118.

{조성배·하성호 (2020). 구조적 토픽 모델링을 활용한 공공데이터 수요 분석. 〈정보화 연구〉, 17권 2호, 103-118.}

Cho, S., Shin, S., Kang, D. (2018). "A Study on the Research Trends on Open Innovation using Topic Modeling." *Informatization Policy, 25*(3), 52-74.

{조성배·신신애·강동석 (2018). 토픽 모델링을 이용한 개방형 혁신 연구동향 분석 및 정책 방향 모색. 〈정보화정책〉, 25권 3호, 52-74.}

Choi, J., Kim, H. & Im, N. (2011). "Keyword Network Analysis for Technology Forecasting." *Journal of Intelligence and Information Systems, 17*(4), 227-240.

{최진호·김희수·임남규 (2011). 기술예측을 위한 특허 키워드 네트워크 분석. 〈지능정보연구〉, 17권 4호, 227-240.}

Dawes, S., Vidiasova, L. & Parkhimovich, O. (2016). "Planning and Designing Open Government Data Programs: An Ecosystem Approach." *Government Information Quarterly, 33*(1), 15-27.

Han, H., Hwang, S., Lee, J. & Oh, H. (2020). "Analysis of Current Status and Improvement Plans of the User Service in Open Data Portal - Focusing on Citizen Participation Data Portal." *Journal of Korean Library and Information Science Society, 51*(1), 255-278.

{한희정·황성욱·이정민·오효정 (2020). 공공데이터포털 이용자 서비스 현황 분석 및 개선방안 - 시민참여형 데이터포털을 중심으로. 〈한국도서관·정보학회지〉, 51권 1호, 255-278.}

Higuchi, K. (2016). *KH Coder 3 manual*. Japan: Ritsumeikan University

Iem, Y., Shim, T. & Lee, S. (2015). "The Study of University Leadership Competencies Compared with University Students' Life Competencies through Co-Word Analysis." *Korean Journal of Family Welfare, 20*(1), 133-151.

{임윤서·심태은·이송이 (2015). 동시출현단어 분석을 통한 대학생의 생애역량과 대학의 리더십역량 비교 연구. 〈가족복지학〉, 20권 1호, 133-151.}

Kubler, S., Robert, J., Neumaier, S., Umbrich, J., & Le Traon, Y. (2018). "Comparison of Metadata Quality in Open Data Portals using the Analytic Hierarchy Process." *Government Information Quarterly, 35*(1), 13-29.

Lee, J. & Park, J. (2019). "An Approach to Constructing a Knowledge Graph Based on Korean Open-Government Data." *Applied Sciences, 9*(19), 1-12.

Lourenço, R. (2015). "An Analysis of Open Government Portals: A Perspective of Transparency for Accountability." *Government Information Quarterly, 32*(3), 323-332.

Ministry of Interior and Safety (2013). *Act on Promotion of the Provision and Use of Open Government Data*.

{행정안전부 (2013). "공공데이터 제공 및 이용 활성화에 관한 법률", 2020년 8월 11일.}

Moon, H. & Lee, K. (2018). "A Study on Public Policy Through Semantic Network Analysis of Public Data related News in Korea." *Journal of Broadcasting Engineering, 23*(4), 536-548

{문혜정·이경서 (2018). 국내 공공데이터 관련 뉴스 의미망 분석을 통한 공공정책 연구. 〈방송공학회논문지〉, 23권 4호, 536-548.}

National Information Society Agency (2017). *The Collection of Informational Statistics in 2017*. Daegu: National Information Society Agency

{한국정보화진흥원 (2017). 〈2017년 정보화통계집〉 대구: 한국정보화진흥원.}

Ohemeng, F. & Ofosu-Adarkwa, K. (2015). "One Way Traffic: The Open Data Initiative Project and the Need for an Effective Demand Side Initiative in Ghana.", *Government Information Quarterly, 32*(4), 419-428.

Open Data Strategy Council (2019). "The 3rd ('20~'22) Basic Plan for Activating Open Data Provision and Utilization." January 10.

{공공데이터전략위원회 (2019). "제3차 공공데이터 제공 및 이용활성화 기본계획." 2020년 1월 10일.}

Palshikar, G. (2007). *Keyword Extraction from a Single Document Using Centrality Measures*. Paper presented at International Conference on Pattern Recognition and Machine Intelligence, December 18-22.

Park, J., Kim, N. & Han, E. (2018). "Analysis of Trends in Science and Technology using Keyword Network Analysis." *Journal of the Korea Industrial Information System Research, 23*(2), 63-73.

{박주섭·김나랑·한은정 (2018). 키워드 네트워크 분석을 활용한 과학기술동향 분석. 〈한국산업정보학회논문지〉, 23권 2호, 63-73.}

Rha, J. (2020). "A Study on the Research Trends in Supply Chain Management in Korea using Network Text Analysis." *Journal of the Korea Industrial Information System Research, 25*(1), 41-53.

{나진성 (2018). 공급사슬관리 국내연구동향 분석: 네트워크 분

석을 활용하여. 〈한국산업정보학회논문지〉, 25권 1호, 41-53.}

Ruijer, E., Grimmelikhuijsen, S. & Meijer, A. (2017). "Open Data for Democracy: Developing a Theoretical Framework for Open Data Use." *Government Information Quarterly, 34*(1), 45-52.

Ruijer, E. & Meijer, A. (2020). "Open Government Data as an Innovation Process: Lessons from a Living Lab Experiment." *Public Performance & Management Review, 43*(3), 613-635.

Seo, H. (2017). "An Empirical Study on Open Government Data: Focusing on ODB and OUR Index." *Information Policy, 24*(1), 48-78.

{서형준 (2017). 공공데이터 개방에 관한 실증 연구: ODB와 OUR Index를 중심으로. 〈정보화정책〉, 24권 1호, 48-78.}

Seo, H. & Myeong, S. (2014). "Policy Alternatives for User-oriented Public Data Utilization - Focusing on ICT Managers' Perception in Private Sector." *The Korean Association for Regional Information Society, 17*(3), 61-86.

{서형준·명승환 (2014). 수요자 중심의 공공데이터 민간 활용 방안. 〈한국지역정보화학회지〉, 17권 3호, 61-86.}

Suh, B. & Shin, S. (2017). "A Study on the Research Trends on Domestic Platform Government using Topic Modeling." *Informatization Policy, 24*(3), 3-26.

{서병조·신선영 (2017). 토픽 모델링을 활용한 한국의 플랫폼정부 연구동향 분석. 〈정보화정책〉, 24권 3호, 3-26}

Tammisto, Y. & Lindman, J. (2012). *Definition of Open Data Services in Software Business*. Paper presented at International Conference of Software Business, June 18-20.

Thorsby, J., Stowers, G., Wolslegel, K. & Tumbuan E. (2017). "Understanding the Content and Features of Open Data Portals in American Cities." *Government Information Quarterly, 34*(1), 53-61.

Unpublished: Google. "Translation API." https://cloud.google.com/translate/ ?hl=ko. (Retrieved on August 10, 2020).

Unpublished: Princeton University. "WordNet: A Lexical Database for English." https://wordnet.princeton.edu/download. (Retrieved on August 10, 2020)

Veljković, N., Bogdanović-Dinić, S., & Stoimenov, L. (2014). "Benchmarking Open Government: An Open Data Perspective." *Government Information Quarterly, 31*(2), 278-290.

Wang, V. & Shepherd, D. (2020). "Exploring the Extent of Openness of Open Government Data-a Critique of Open Government Datasets in the UK." *Government Information Quarterly, 37*(1), 1-10.

Worthy, B. (2015). "The Impact of Open Data in the UK: Complex, Unpredictable, and Political." *Public Administration, 93*(3), 788-805.

Yun, S. & Hyun, J. (2019). "An Analysis of Open Data Policy in Korea: Focused on National Core Data in Open Data Portal." *Korean Public Management Review, 33*(1), 219-247.

{윤상오·현지우 (2019). 공공데이터 개방정책의 실태분석 및 개선방안에 관한 연구: 공공데이터 포털의 국가중점데이터 개방 사례를 중심으로. 〈한국공공관리학보〉, 33권 1호, 219-247.}

Zeleti, F., Ojo, A. & Curry, E. (2016). "Exploring the Economic Value of Open Government Data." *Government Information Quarterly, 33*(3), 535-551.