Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography
Vol. 35, No. 6, 461-470, 2017
https://doi.org/10.7848/ksgpc.2017.35.6.461

ISSN 1598-4850(Print)
ISSN 2288-260X(Online)
Original article

사회 연결망 분석을 활용한 공공데이터 간 연관성에 관한 연구

A Study on the Linkability of Public Information Using Social Network Analysis

정다운[1] · 이미숙[2] · 신동빈[3]
Jeong, Da Woon · Yi, Mi Sook · Shin, Dong Bin

Abstract

In Korea, starting with the Government 3.0 Policy, the utilization of public data as an important driving force to promote economic growth has been highlighted as a major issue. However Korea is currently only able to open and provide accumulated data stored in the public domain. To resolve this issue, we need to not only open and provide public information, but also to create new information by linking the data and developing related services. Thus, this study analyzes the linkability of public information and provides lists of the linkable public data. In order to do this, we first have performed preconditioning processes on the accessibility and workability of the data. Next, we have deduced the major keywords in public data through analyzing the morphemes, and then the core keywords (Top 10) and their linkable keyword lists through an analysis of social networks. Based on the outcome of this study, a subsequent study will deduce new information by linking the public data and creating various services and information contents. Furthermore, not only conceptual but also practical linking measures need to be created, and a related law must be prepared.

Keywords: Public Data, Social Network Analysis, Keyword, Linkability

초 록

한국은 정부 3.0 정책을 기조로 하여 경제 성장을 증진하기 위한 주요 추진 동력으로써 공공데이터의 활용이 주요 이슈로 부각되고 있다. 그러나 한국정부는 현재 공공 영역에 축적되어 있는 데이터의 공개나 제공 수준에 머무르고 있다. 따라서 단순 공공데이터뿐만 아닌 공공데이터 간의 연계를 통한 새로운 정보를 창출하고, 관련 서비스의 개발 등이 요구되고 있다. 이에 본 연구는 공공데이터 목록을 수집 및 정제하고, 사회 연결망 분석을 통해 핵심 주제별 연관성이 높은 공공데이터 정보 목록을 도출하였다. 이를 위해서 첫째, 수집한 공공데이터 목록을 지자체 담당자를 대상으로 설문조사를 수행하였다. 이를 통해 접근 용이성 측면과 가공 용이성 측면에서 전처리 과정을 수행하여 불필요한 공공데이터를 정제하였다. 다음으로 개념적인 차원에서의 공공데이터 간 연관성을 분석하기 위해서 형태소 분석을 통해 공공데이터의 대표 키워드를 도출하였다. 이후 사회 연결망 분석을 활용하여 핵심 키워드(상위 10개) 및 연관성이 높은 공공데이터 목록을 도출하였다. 본 연구결과를 바탕으로, 향후에는 공공데이터 간 연계를 통해 융·복합된 새로운 정보를 기반으로 다양한 스마트시티 서비스를 창출할 수 있을 것으로 전망된다. 또한, 이를 위해서는 개념적 연계뿐만 아니라 실질적인 연계 방안이 도출되어야 할 것이며, 이에 따른 법·제도적 정비도 필요할 것으로 사료된다.

핵심어 : 공공데이터, 사회 연결망 분석, 키워드, 연관성

Received 2017. 10. 24, Revised 2017. 11. 16, Accepted 2017. 12. 18

1) Graduate School of Urban Information Engineering, Anyang University (E-mail: daun5342@naver.com)
2) Corresponding Author, Member, Dept. of Urban Information Engineering, Anyang University (E-mail: mslee0414@anyang.ac.kr)
3) Member, Dept. of Urban Information Engineering, Anyang University (E-mail: dbshin@anyang.ac.kr)

1. 서 론

한국사회는 IT가 일상화되면서 시민사회와 산업계로부터 생활·문화·지식 콘텐츠에 대한 공개 요구가 증가하고 있으며, 원천데이터로 공공데이터에 대한 활용 요구도 함께 늘어나고 있다. 이에 따라 현재 한국의 공공기관은 사회, 경제, 지리 등 여러 분야의 활동 영역에서 폭넓은 정보를 수집, 생산, 재생산 및 분배하여 활용 중에 있다. 본래 공공기관은 공적 기능을 위해 많은 비용을 들여 정보를 생성, 수집, 관리하며, 그 목적을 달성하면 해당 정보는 폐기되거나 방치되는 것이 일반적이었다.

그러나 최근 들어 인터넷과 디지털 기술, 공공기관 정보를 민간에서 어플리케이션 개발 등에 바로 활용 가능하도록 하는 공유 프로그램의 표준 인터페이스인 OpenAPI와 매시업 등이 공공데이터의 새로운 활용 가능성을 열어주고 있다. 이는 공공데이터의 재활용을 통해 새로운 부가가치를 창출함으로써 다양한 지식정보 서비스로 활용될 수 있는 가능성을 열어주고 있다는 것을 의미한다.

이와 같은 맥락으로 공공데이터의 개방과 활용은 한국 정부와 시민 간의 소통 활성화를 통한 열린 정부 구현, 국민의 정보채널 선택권 확장, 대국민 행정서비스의 질 제고, 협업 구현 등에 있어서 중요한 의의를 지니고 있다. 또한 4차 산업혁명에 대응하기 위해서 국내 스마트시티 및 스마트시티 서비스에서의 공공데이터 활용은 해당 분야에 주요 이슈로 자리잡고 있다. 이러한 배경에서 본 연구는 한국의 공공데이터포털에서 제공하는 국가행정기관과 지방자치단체의 공공데이터 목록을 대상으로 공공데이터 간의 융·복합을 위한 연관성을 분석하고자 한다. 분석결과를 토대로 주요 분야별 연관성이 높은 공공데이터 목록(안)을 제시하는 것에 연구의 목적을 두고 있다.

2. 선행연구 검토

본 연구와 관련하여 정보 공개 및 활용, 사회 연결망 분석(Social Network Analysis) 등 2개 측면에서의 선행연구들을 살펴보았다.

우선적으로 정보 공개 및 활용과 관련하여, 공공데이터 민간 활용에 관한 몇 가지 쟁점 연구에서는 주요 법적 쟁점으로 제공대상, 제공과 제공중단, 이용 비용의 산정과 부과상에서의 적정성, 분쟁조정 측면에서 공공데이터 활용의 주요 쟁점을 보고 있다(Kim *et al.,* 2014). 다음으로 수요자 중심의 공공 데이터 민간 활용 방안 연구에서는 공공데이터 활용 장애 요인으로 정부의 공공 정보 제공 환경 및 기반 측면, 기술적 환경, 정보 환경, 사회·경제적 환경 분야에서 측정하여, 가장 큰 장애요인으로 데이터 품질의 한계를 제시했다(Seo and Myeong, 2014). 정보 공개 및 활용과 관련한 이와 같은 선행연구는 공공데이터 활용의 장애요인을 분석·제시하였다는데 의의가 있지만, 주로 법제도적 논의에 그치고 있다는 점에서 한계가 있다. 본 연구에서는 공공데이터 간의 연계를 통해 새로운 정보를 창출하기 위하여 공공데이터간의 연관성을 분석하였다는 점에서 선행연구와는 차별성이 있다.

다음으로 사회 연결망 분석과 관련하여, 국내 연구로는 한국 간호학 연구 주제의 사회연결망 분석을 통해 중심연구주제 파악과 새로운 주제의 변화 추이를 3년 주기로 분석하였고, 연구주제가 간호 실무와 연계됨을 보여주었다(Jang *et al.,* 2012). 또한, 암유전체에 대한 국내외 연구주제를 시계열적으로 비교분석하여 연구주제 비교를 통해 연구수준의 확인과 향후 활발하게 연구될 주제와 연구기관의 협동 관계를 확인하였다(Lee *et al.,* 2011). 이와 같은 연구는 사회 연결망 분석을 활용하여 중심연구주제를 파악하였다는데 의의가 있으며, 본 연구에서도 핵심 주제별 연관성이 높은 공공데이터 정보 목록을 도출하는데 이러한 사회 연결망 분석을 활용할 수 있을 것이다.

이와 같이 기존 선행연구는 주로 공공기관 간 공공데이터의 공유 및 활용에 관한 내용이 주를 이루고 있으며, 공공데이터 간 연관성에 대한 선행연구는 현재 미비한 상태이다. 이에 본 연구에서는 공공데이터 간 개념적 연관성 분석을 사회 연결망 분석을 통해 수행하고, 이에 따라 공공데이터 간의 연관성이 높은 정보 목록을 도출하고자 한다.

3. 분석의 설계

3.1 분석의 대상

본 연구에서는 공공데이터포털에서 제공하는 공공데이터(2016년 기준)를 대상으로 분석을 수행하였다. 공공데이터 기관은 중앙정부(17부, 5처, 16청, 2016년 기준)와 스마트시티를 추진하는 지자체(안양시, 인천광역시 등)를 대상으로 하였으며, 공공데이터 개수는 데이터를 기준으로 총 5,660개로 나타났다(중앙정부 1,908개, 지자체 3,752개). 공공데이터 담당 기관 설정과 관련하여, 본 연구의 결과인 개념적 차원에서의 공공데이터 간 연관성 분석결과를 토대로 위치정보 기반의 스마트시티 서비스 기획에 반영하고자 실제 스마트시티 추진 지자체를 대상으로 공공데이터를 수집하였다.

3.2 분석의 절차 및 방법

본 연구는 Fig. 1과 같은 절차로 진행하였다.

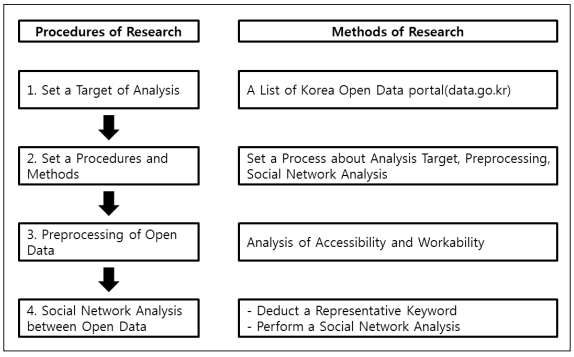

Fig. 1. Procedures and methods of research

먼저, 본 연구에서 수집한 공공데이터 목록에 대해 공공데이터 및 스마트시티 관련 지자체 담당자의 설문조사를 통해 공공데이터의 접근 용이성 측면과 가공 용이성 측면에서 전처리 과정을 수행하였다. 마지막으로 형태소 분석을 통해 공공데이터의 대표 키워드(분야)를 도출하고, 사회 연결망 분석을 토대로 공공데이터 간 연관성이 높은 데이터 목록을 도출하였다.

4. 분석 결과

4.1 공공데이터의 전처리 과정

공공데이터의 전처리 과정은 접근 용이성과 가공 용이성 측면에서 진행하였다. 접근 용이성은 사용자가 해당 시스템(정보)에 접근을 위해 공공데이터를 어느 방식으로 제공 받는지를 의미한다. 가공 용이성은 공공데이터 간 연계 또는 활용과 관련하여 사용자가 손쉽게 공공데이터 형식을 가공 할 수 있는지를 의미한다.

이에 따라 본 연구에서는 공공데이터의 제공 방식과 데이터 구축 형태에 대해 공공데이터 관련 지자체 담당자를 대상으로 설문조사를 수행하였으며, 그 결과는 다음과 같다. 지자체 담당자 설문은 공공데이터 관련 지자체 공무원(만 19세 이상 성인 남녀)을 대상(리커트 5점 척도 기준)으로 총 50부를 배포하였으며, 32부를 회수(회수율 62%)하였다. 설문조사 내용은 공공데이터 제공 방식(6개)과 20개의 데이터 형태로 구분하여 수행하였다. 데이터 형태의 경우에는 중앙정부 및 지자체에서 수집한 공공데이터 목록의 형식들을 기준으로 정리하였다.

4.1.1 공공데이터의 접근 용이성 관련 설문조사 결과

공공데이터 제공 방식은 다운로드, Link, OpenAPI, 그리드, 대행판매, 활용신청 등으로 구분된다. 이 중에서 다운로드 및 Link 등의 혼용된 방식으로 공공데이터를 제공하는 경우도 있다. 공공데이터 제공 방식별 개수는 Table 1과 같다. 설문조사 결과, 다운로드(합계 127점, 평균 3.97점) 방식이 가장 접근 용이성이 높은 것으로 나타났으며, 다음으로 OpenAPI (합계 115점, 평균 3.60점) 등의 순으로 결과가 도출되었다. Link와 대행판매, 활용신청 방식은 모두 평균 1점 대에 해당하는 수치를 보였으며, 해당 결과에 따라 Link 등 기타 홈페이지로 2번 이상 이동되는 경우와 데이터 수집을 위한 대행판매, 활용신청 등은 공공데이터 간 연계를 위한 접근 용이성이 낮은 것을 알 수 있었다. 이에 Link, 대행판매, 활용신청 등 접근 용이성이 상대적으로 낮은 항목에 해당하는 공공데이터를 분석 대상에서 제외하였다.

Table 1. Analysis result of the accessibility by each provision method

Provision Method	Amount	Accessibility
Download	3,961	High
Link	1,433	Low
OpenAPI	57	High
Sales	2	Low
Application	3	Low
Null	204	-

4.1.2 공공데이터의 가공 용이성 관련 설문조사 결과

공공데이터 구축 형태는 DB, DOC/DOCX, DXF 등으로 구분된다. 공공데이터 제공 방식과 마찬가지로 공공데이터 목록 중에서도 두 가지 이상의 방식을 혼용하여 제공하는 경우도 존재한다. 공공데이터의 구축 형태별 개수는 Table 2와 같다. 설문조사 결과, XLS/XLSX(합계 135점, 평균 4.22점) 형태가 가장 가공 용이성이 높은 것을 알 수 있었다. 이는 주로 분석을 목적으로 한 텍스트 및 프로그램 파일 등의 데이터베이스 파일로 구축된 항목들이 공공데이터 간 연계를 위한 가공 용이성이 가장 높은 것으로 집계된 것을 알 수 있다. 또한 프로그램 데이터 포맷 중에서는 SHP 형식만이 높은 수치를 보였으며, 영상/이미지 파일 중에서는 JPG 파일만이 가공 용이성이 높은 것을 알 수 있다. 20개의 데이터 형태 모두 평

Table 2. Analysis result of the workability by each format

Format	Amount	Accessibility	Format	Amount	Accessibility
DB	4	High	SGM	2	High
DOC/DOCX	23	Low	SHP	13	High
DXF	1	High	TXT	103	High
EXE	14	High	XLS/XLSX/CSV	3,516	High
HTML	93	High	XML	17	High
HWP	861	High	Other	193	Low
JSON	1	High	Video	21	Low
JSP	16	High	Image	61	Low
PDF	81	High	Website	597	Low
PPT/PPTX	3	High	Null	40	-

균 2.47점 이상이었으며, 공공데이터의 가공 용이성이 '매우 낮은' 항목은 없었다. 즉, 20개의 데이터 형태 모두 공공데이터 간 연계 시 가공이 용이한 데이터 형태로 볼 수 있으나, 동영상, 이미지, 홈페이지, 기타 등 가공 용이성이 상대적으로 낮은 항목에 해당하는 공공데이터를 분석에서 제외하였다.

결과를 종합하면, 전처리 후 공공데이터의 접근 용이성과 가공 용이성 등 두 개의 기준을 충족한 공공데이터는 총 3,985개로 나타났다.

4.2 사회 연결망 분석

4.2.1 공공데이터의 대표 키워드 도출을 위한 형태소 분석결과

앞서 전처리 과정을 거친 공공데이터 목록을 대상으로 분야별 대표 키워드를 도출하기 위해서 형태소 분석을 수행하였다. 형태소 분석은 단어를 구성하는 각각의 형태소를 인식하고 불규칙 활용이나 축약, 탈락 현상이 일어난 형태소는 원형을 복원하는 과정을 말한다. 이를 위해서 국립국어원에서 배

Table 3. Frequency of the major keywords in public information

Keyword	Frequency	Keyword	Frequency	Keyword	Frequency
Status	1,689	Crime	120	Population	111
Arrest	108	Car	80	Foreigner	77
Water Quality	69	Industry	66	Resident	65
Medicine	58	Employment	55	Training	53
Statistics	51	Collection	42	Research	41
Disabled	40	Railroad	40	Age	39
Region	38	Elderly	37	Youth	37
Architecture	35	Computing	34	FTA	33
⋮	⋮	⋮	⋮	⋮	⋮

포하는 오픈 소프트웨어인 지능형 형태소 분석기를 활용하여, 공공데이터 목록 총 3,985개의 데이터 명칭을 품사 형태로 분리하였다. 또한 형태소 분석 과정 중에서 공공데이터의 제목이 없거나 오타 등이 있는 경우에는 불완전 용어로 간주하여 해당 목록에서 제외하였다. 그 결과, Table 3과 같이 총 1,195개의 대표 키워드가 도출되었다.

4.2.2 대표 키워드에 따른 사회 연결망 분석결과

상위 빈도를 갖는 200개의 대표 키워드를 대상으로 사회 연결망 분석을 수행하였으며, 그 결과는 Fig. 2와 같다.

사회 연결망 분석은 근본적으로 네트워크에서 노드 간의 연관관계에 따라 노드의 역할이 네트워크에 어떤 역할을 담당하거나 영향을 미치는 것을 파악하는 것이 중요하다. 네트워크의 중심성 분석방법은 각각의 행위자를 표현하는 노드와 상호 연관관계를 의미하는 링크에 네트워크를 구성하게 되어 있으며, 이들의 상관관계를 규명하는 것이 중요하다. 특히 어떤 노드가 허브 역할을 하는지, 어떤 노드가 중계자의 역할을 하는지 등의 중심에 대한 해석이 가장 중요하며, 전체 네트워크에서 각각의 노드들에 대한 역할, 위치, 특성, 영향력을 파악하는 것이 중요하다. 또한 네트워크 내에서 노드들의 상대적 위치나 절대적 위치를 파악하는데 유용하게 활용할 수 있다. 네트워크 관련 분석방법 중 본 연구에서 활용한 중심성 분석은 연결 중심성, 매개 중심성, 근접 중심성으로 세분화 된다. 이러한 중심성은 영향력이라는 개념으로 해석되기도 하

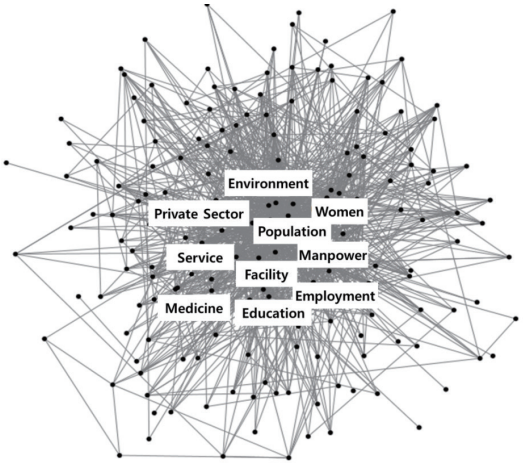

Fig. 2. SNA result of the core keywords (top 10) in public information

며, 일반적으로 가장 많이 사용되는 분석 기법 가운데 하나이다. 연결 중심성은 하나의 노드에 직접적으로 연결되어 있는 다른 노드의 수를 말하며, 매개 중심성은 한 노드가 연결망 내 다른 노드 사이에 위치하는 정도를 의미한다. 근접 중심성은 한 노드로부터 다른 노드에 도달하기 위한 최소 단계의 거리를 뜻한다.

분석결과, 공공데이터 관련 연결 중심성을 기준으로 높은

Table 4. Social network index of the core keywords (Top 10) in public information

Order	Core Keyword	Frequency	Degree	Betweenness	Closeness
1	Facility	57	0.016	798.193	0.003
2	Manpower	32	0.013	111.382	0.003
3	Service	24	0.011	304.009	0.003
4	Women	24	0.011	48.699	0.003
5	Population	23	0.010	54.867	0.003
6	Education	28	0.010	210.578	0.003
7	Private Sector	19	0.009	28.742	0.003
8	Employment	20	0.009	34.503	0.003
9	Environment	16	0.009	6.414	0.003
10	Medicine	16	0.008	48.087	0.003
⋮	⋮	⋮	⋮	⋮	⋮

수치를 보이는 대표 키워드 상위 10종을 핵심 키워드로 정의하였다. 이를 살펴보면 시설, 인력, 서비스, 여성, 인구, 교육, 민간, 고용, 환경, 의료 등 다양한 분야에서의 공공데이터별 대표 키워드가 도출되는 것을 확인할 수 있었다. 연결 중심성이 높다는 것은 키워드 간 연관성이 높은 공공데이터 목록에서 다른 대표 키워드와 함께 많이 사용된 키워드를 의미하는 것으로, 이는 다른 공공데이터와의 융합 가능성이 높은 핵심 키워드로 볼 수 있다. 매개 중심성을 분석한 결과는 Table 4와 같이 연결 중심성과 마찬가지로 시설, 서비스, 교육, 인력 등의 키워드가 높은 것을 확인할 수 있었다. 그러나 연결 중심성 결과와 비교했을 때, 수질, 시정, 성별, 특허, 외국인, 범죄 등 새로운 키워드를 확인할 수 있으며, 이는 기존 핵심 키워드와 융합하기 위한 연결고리 역할이 가능하다. 마지막으로 근접 중심성은 다른 키워드에 접근하기 위해 영향력이 높은 키워드를 말하며, 위의 분석 결과에서는 대표 키워드별 근접 중심성이 유사하게 나타난 것을 알 수 있었다.

다음으로 핵심 키워드별 연관성이 높은 정보 목록을 도출하기 위해서, 핵심 키워드를 기준으로 매개 중심성이 높은 키워드 목록을 살펴보았다. 매개 중심성은 앞서 명시한 핵심 키워드와의 융합 관련 연결고리 역할을 하는 키워드로, 매개 중심성이 높은 키워드를 연계 정도가 높은 키워드로 정의하였다. 해당 분석과 관련하여, 근접 중심성 지수는 키워드 간의 격차가 크지 않기 때문에 분석에 포함하지 않았다. 시설의 경우 Fig 3.에서 보는 바와 같이 서비스, 수질, 생활, 여성, 의료, 문화, 민간, 하수, 안전, 주민 키워드의 매개 중심성이 기타 키워드에 비해 높게 나타났으며, 이는 시설 키워드와의 연관성이 높다는 것을 의미한다. Table 5를 보면, 해당 키워드 중에서도 시설과 연관성이 높은 대표 키워드는 서비스로 나타났다.

Fig. 3. Network of linkable information for the keyword 'facility'

인력의 경우 성별, 여성, 의료, 고용, 민간, 보건, 직업, 학교, 도서, 진료 키워드의 매개 중심성이 기타 키워드에 비해 높게

Table 5. List of linkable information for the keyword 'facility'

Order	Keyword	Frequency	Betweenness Centrality	Closeness Centrality	Degree Centrality
1	Service	24	304.009	0.003	0.011
2	Water Quality	8	203.148	0.002	0.003
3	Life	13	51.916	0.002	0.005
4	Women	24	48.699	0.003	0.011
5	Medicine	16	48.087	0.003	0.008
6	Culture	18	45.536	0.003	0.006
7	Private Sector	19	28.742	0.003	0.009
8	Sewage	17	25.343	0.003	0.007
9	Safety	15	21.560	0.003	0.007
10	Resident	10	17.592	0.003	0.005

나타났으며, 인력과 연관성이 높은 대표 키워드는 성별로 나타났다. 서비스의 경우 시설, 교육, 여성, 상표, 장애인, 아동, 지식, 기술, 지역, 연구 키워드의 매개 중심성이 기타 키워드에 비해 높게 나타났으며, 서비스와 연관성이 높은 대표 키워드는 시설로 나타났다. 여성의 경우 시설, 서비스, 교육, 인력, 민간, 직업, 가정, 장애인, 기업, 농업 키워드의 매개 중심성이 기타 키워드에 비해 높게 나타났으며, 여성과 연관성이 높은 대표 키워드는 시설로 나타났다. 인구의 경우 교육, 성별, 외국인, 국적, 하수, 세대, 주민, 경제, 건강, 연령 키워드의 매개 중심성이 기타 키워드에 비해 높게 나타났으며, 인구와 연관성이 높은 대표 키워드는 교육으로 나타났다. 교육의 경우 서비스, 범죄, 인구, 여성, FTA, 범죄자, 보건, 안전, 직업, 학교 키워드의 매개 중심성이 기타 키워드에 비해 높게 나타났으며, 교육과 연관성이 높은 대표 키워드는 서비스로 나타났다. 민간의 경우 시설, 인력, 여성, 급수, 체육, 비상, 기술, 연구, 기업, 과학 키워드의 매개 중심성이 기타 키워드에 비해 높게 나타났으며, 민간과 연관성이 높은 대표 키워드는 시설로 나타났다. 고용의 경우 성별, 인력, 안전, 수급, 연령, 노동, 석면, 기술, 지역, 산업 키워드의 매개 중심성이 기타 키워드에 비해 높게 나타났으며, 고용과 연관성이 높은 대표 키워드는 성별로 나타났다. 환경의 경우 시설, 단속, 위생, 보건, 징수, 대기, 배출, 사업장, 청소년, 조사 키워드의 매개 중심성이 기타 키워드에 비해 높게 나타났으며, 환경과 연관성이 높은 대표 키워드는 시설로 나타났다. 의료의 경우 시설, 인력, 처분, 단속, 보건, 수급, 병원, 약국, 노인, 예방 키워드의 매개 중심성이 기타 키워드에 비해 높게 나타났으며, 의료와 연관성이 높은 대표 키워드는 시설로 나타났다.

이와 같이 핵심 키워드별 연관성이 높은 키워드를 종합하면, Table 6과 같다. 핵심 키워드와 연관성이 높은 키워드를 본래의 공공데이터로 전환하여 살펴보면, 해당 결과를 토대로 공공데이터 간 연관성이 높은 정보 목록을 확인할 수 있다.

Table 6. List of linkable data in public information

Order	Field	Linkable Keyword	Public Information (Typical)	Order	Field	Linkable Keyword	Public Information (Typical)
1	Facility	Service	Information Search Service	2	Human Resources	Gender	Age Info. by Gender
		Water Quality	Water Quality Inspection Status			Women	Female Household
		Life	Public Assistance Recipient Status			Medicine	Medical Facility Status
		Women	Female Household			Employment	Employment Trend
		Medicine	Medical Facility Status			Private Sector	Private Facility Status
		Culture	Cultural Property Status			Public Health	Health Center Status
		Private Sector	Private Facility Status.			Occupation	Employee Status by Occupation
		Sewage	Sewage Disposal Plant Status			School	School Status
		Safety	Road Facility Safety Inspection			Book	Books Status
		Resident	Registered Population Statistics			Medical Treatment	Clinic Status

3	Service	Facility	Facility Status	4	Women	Facility	Facility Status Info
		Education	Educational Institution Status			Service	Information Search Service
		Women	Female Household			Education	Educational Institution Status
		Trademark	Integrated Trademark History			Manpower	Manpower by Degree/ Gender
		Disabled	Registered Disabled Status			Private Sector	Private Facility Status
		Children	Child Center Status			Occupation	Employee Status by Occupation
		Knowledge	Regional Intellectual Property Center Status			Family	Group Home Status
		Technology	National Information Technology Standard			Disabled	Registered Disabled Status
		Region	Usage District			Business	Small & Medium-Sized Business Statistics
		Research	National R&D Project Investment			Agriculture	Agriculture Status
5	Population	Education	Educational Institution Status	6	Education	Service	Information Search Service
		Gender	Age Info. by Gender			Crime	Crime Place
		Foreigner	Foreign Resident Status			Population	Population Status
		Nationality	Population Info. by Nationality			Women	Female Household
		Sewage	Sewage Disposal Plant Status			FTA	FTA Agreement Tax Rate
		Household	Household Status			Criminal	Criminal Execution Status
		Resident	Registered Population Statistics			PublicHealth	Health Center Status
		Economy	Economically Active Population			Safety	Safety Inspection Result
		Health	Health-Insured Population Status			Occupa tion	Employee Status by Occupation
		Age	Population Info. by Age			School	School Status Info

7	Private Sector	Facility	Facility Status Info	8	Employment	Gender	Age Info. by Gender
		Manpower	Manpower Info. by Degree/Gender			Manpower	Manpower by Degree/ Gender
		Women	Female Household			Safety	Road Facility Safety Inspection
		Water Supply	Water Supply Facility			Supply and Demand	Public Assistance Recipient Status
		Sports	Sports Facility Status			Age	Population Info. by Age
		Emergency	Emergency Shelter Status			Labor	Labor Transfer Status
		Technology	National Information Technology Standard			Asbestos	Asbestos Inspection Subject Building Status
		Research	National R&D Project Investment			Technology	National Information Technology Standard
		Business	Small & Medium-sized Business Statistics			Region	Usage District
		Science	Interest Index by Science Technology Field			Industry	Technology Export Progress by Industry
9	Environment	Facility	Facility Status	10	Medicines	Facility	Facility Status
		Control	CCTV Status			Manpower	Manpower Info. by Degree/Gender
		Sanitation	Public Sanitary Institutions Status			Execution	Medical Institution Administrative Execution Status
		Sanitation	Public Sanitary Facility Status			Control	CCTV Status
		Collection	Collection Status			Public Health	Health Center Status
		Atmosphere	Air Pollution Measurement			Supply and Demand	Public Assistance Recipient Status
		Emission	Discharging Facility Status			Hospital	Hospital Status
		Establishment	Dust Scattering Establishment Status			Pharmacy	Pharmacy Status
		Youth	Youth Facility Status			Elderly	Senior Population Status
		Research	Business Status Research			Prevention	Vaccination Status

5. 결 론

본 연구는 한국의 공공데이터포털에서 공개 및 제공하는 공공데이터 목록을 토대로 공공데이터간의 사회 연결망 분석을 통해 연관성이 높은 공공데이터 목록을 도출하였다.

이에 일환으로 본 연구에서는 먼저 중앙정부 및 지자체에서 제공하는 공공데이터의 전처리 과정을 수행하였다. 전처리 과정은 설문조사 결과를 기준으로 공공데이터의 제공 방식을 고려한 접근 용이성 측면, 그리고 공공데이터의 구축 형태를 고려한 가공 용이성 측면으로 구분하여 분석을 수행하였다. 그 결과, 공공데이터의 제공방식 및 데이터 유형에 따라 전처리 과정을 거친 공공데이터 개수는 총 3,780개로 도출되었다. 다음으로 전처리 과정을 수행한 공공데이터를 대상으로 연관성이 높은 정보 목록을 도출하기 위하여, 우선적으로 형태소 분석을 통한 공공데이터의 대표 키워드를 도출하였으며, 총 1,195개의 대표 키워드가 도출되었다. 이후 상위 빈도를 갖는 200개의 대표 키워드를 대상으로 사회 연결망 분석을 수행하였으며 그 결과, 공공데이터 관련 연결 중심성을 기준으로 높은 수치를 보이는 대표 키워드 상위 10개를 핵심 키워드로 정의하였다. 상위 10개의 핵심 키워드는 시설, 인력, 서비스, 여성, 인구, 교육, 민간, 고용, 환경, 의료 등 다양한 분야에서의 핵심 키워드가 도출된 것을 알 수 있었다. 다음으로 핵심 키워드와의 연계를 위하여 융합 관련 연결고리 역할을 수행하는 매개 중심성을 토대로 핵심 키워드별 각 10개의 연관성이 높은 키워드를 도출하였다. 마지막으로 핵심 키워드와 연관성이 높은 키워드를 본래의 공공데이터로 전환하여, 공공데이터간 연관성이 높은 목록을 도출하였다.

본 연구에서는 개념적인 차원에서의 공공데이터 간 연관성을 목적으로 분석을 수행하였다. 현재 공공데이터포털에서 제공하는 공공데이터는 단일 파일 형식이 주를 이루고 있으며, 표준데이터는 58건(2017년 기준)에 불과하다. 단일 형식으로 제공되는 파일데이터는 데이터베이스 테이블의 개념이 부재하며, 따라서 DB와 각 파일데이터 간 연계는 현실적으로 불가능할 것으로 사료된다. 따라서, 본 연구에서는 단일 데이터 측면에서의 연계를 위한 연관성 분석을 수행하였는다는 점에서 공공데이터포털에서 제공하는 공공데이터의 분류체계와는 차별성이 있다.

또한 다양한 단일 데이터의 종류에 비해 표준데이터의 종류는 현저하게 낮은 실정이며, 무엇보다도 표준데이터의 최신성 및 정확성 측면에서 유지관리의 어려움이 존재한다. 이에 다양한 분야에서의 단일 데이터들을 표준데이터의 속성(컬럼) 항목을 기준 양식으로 하여, 각 기관에서 표준데이터가 업로드 될 수 있는 체계 구축 방안이 마련될 필요가 있다.

향후 연구에서는 공공데이터 간의 개념적 연계뿐만 아닌 실질적인 연계 방안이 제시되어야 한다. 이와 관련해서 다음과 같은 공공데이터 연계를 위한 추가적인 연구가 수반되어야 한다. 우선 해당 분야의 담당자와 시스템 운영자 등의 면담 및 자료 확보를 통한 수요자 측면에서의 요구사항 분석이 수행되어야 한다. 다음으로 구축된 DB를 실질적으로 분석하여 공공데이터 연계 시 문제점을 도출하여 개선사항을 마련하여야 한다.

Acknowledgments

This research was supported by the MOLIT(The Ministry of Land, Infrastructure and Transport) of Korea, under the UPA(Urban Planning & Architecture) research support program supervised by the KAIA(Korea Agency for Infrastructure Technology Advancement) (16AUDP-B070716-04).

References

Jang, H.L., Kang, G.W., Lee, E.J., and Kim, S.R. (2012), Analysis of research subject network in the field of oncogene, *Journal of Korea Technology Innovation Society,* Vol. 15, No. 2, pp. 369-399. (in Korean with English abstract)

Kim, J.W., Lee, D.H., and Bae, S.H. (2014), A study on the current legal issues of the re-use of the public sector data in Korea, *Korean Lawyers Association Journal,* Vol. 63, No. 4, pp. 5-45. (in Korean)

Lee, S.K., Jung, S.W., Kim, H.G., and Yom, Y.H. (2011), A social network analysis of research topics in Korean nursing science, *Journal of Korean Academy of Nursing,* Vol. 41, No. 5, pp. 623-632. (in Korean with English abstract)

Seo, H.J. and Myeong, S.H. (2014), Policy alternatives for user-oriented public data utilization : focusing on ICT managers' perception in private sector, *Journal of Korean Association for Regional Information Society,* Vol. 17, No. 3, pp. 61-86. (in Korean with English abstract)