JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 26 No. 9, pp. 13-25, September 2021
https://doi.org/10.9708/jksci.2021.26.09.013

Development of a Method for Analyzing and Visualizing Concept Hierarchies based on Relational Attributes and its Application on Public Open Datasets

Suk-Hyung Hwang*

*Professor, Dept. of Artificial Intelligence and Software Technology, SunMoon University, A-San, Korea

[Abstract]

In the age of digital innovation based on the Internet, Information and Communication and Artificial Intelligence technologies, huge amounts of datasets are being generated, collected, accumulated, and opened on the web by various public institutions providing useful and public information. In order to analyse, gain useful insights and information from data, Formal Concept Analysis(FCA) has been successfully used for analyzing, classifying, clustering and visualizing data based on the binary relation between objects and attributes in the dataset. In this paper, we present an approach for enhancing the analysis of relational attributes of data within the extended framework of FCA, which is designed to classify, conceptualize and visualize sets of objects described not only by attributes but also by relations between these objects. By using the proposed tool, RCA wizard, several experiments carried out on some public open datasets demonstrate the validity and usability of our approach on generating and visualizing conceptual hierarchies for extracting more useful knowledge from datasets. The proposed approach can be used as an useful tool for effective data analysis, classifying, clustering, visualization and exploration.

▸ **Key words**: Open data, Relational attribute, Data mining, Data visualization, Concept analysis, Conceptual hierarchy

[요 약]

인터넷과 정보통신, 인공지능기술을 기반으로 하는 디지털 혁신 시대를 맞이하면서 거대한 규모의 데이터집합이 발생, 수집, 축적되어, 다양한 공공기관에서 온라인에 오픈하여 유용한 공공정보를 제공하고 있다. 데이터를 분석하여 유용한 통찰력과 정보를 얻기 위하여, 데이터집합에 내재되어 있는 객체와 속성 사이의 이진 관계를 기반으로 데이터를 분석, 분류, 군집화 및 시각화하는 형식개념분석기법이 성공적으로 사용되어 왔다. 본 논문에서는 형식개념분석기법을 확장하여, 객체의 속성뿐만 아니라 객체들 사이의 관련 관계를 기반으로 데이터집합을 분류하고 개념화하여 가시화하기 위한 기법과 지원도구를 제안한다. 일부 공공 오픈 데이터집합을 대상으로 본 논문의 제안기법을 적용하여 몇 가지 실험을 수행한 결과, 데이터집합으로부터 개념 계층구조를 생성하고 시각화하여 보다 유용한 지식을 추출함으로써 제안기법의 타당성과 유용성을 실증하였다. 본 논문에서 제안한 분석기법은 효과적인 데이터분석, 분류, 군집화, 시각화, 정보검색 등을 위한 유용한 도구로 사용될 수 있다.

▸ **주제어**: 오픈데이터, 관계속성, 데이터마이닝, 데이터시각화, 개념분석, 개념계층구조

• First Author: Suk-Hyung Hwang, Corresponding Author: Suk-Hyung Hwang
*Suk-Hyung Hwang (shwang@sunmoon.ac.kr), Dept. of Artificial Intelligence and Software Technology, SunMoon University
• Received: 2021. 05. 27, Revised: 2021. 09. 03, Accepted: 2021. 09. 15.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

데이터는 4차 산업혁명을 견인하는 핵심 자원으로서 국내뿐만 아니라 해외 주요 국가들에서는 정부 차원에서 데이터 산업을 육성하기 위한 다양한 정책 수립과 플랫폼 구축, 공공데이터(Public Open Data) 제공을 위해 노력하고 있으며, 데이터 관리, 데이터분석, 머신러닝 등 데이터 및 데이터 관련 산업이 지속해서 성장하고 있다[1]. 우리나라의 공공데이터 포탈(www.data.go.kr)에서는 교육, 국토관리, 공공행정 등 총 16개 영역으로 분류하여, 956개 기관에서 38,139개의 파일데이터와 6,774개의 오픈 API를 개방하고 있고(2021년 3월 17일 현재)[2], 최근 4년간의 공공데이터 개방 및 활용이 크게 증가하여, 2021년 코로나19의 확산과 함께 국내 데이터 시장은 2024년까지 연평균 15.3%씩 성장하여 약 30조 원 규모에 이를 전망이다[3].

최근에는 인터넷과 정보통신, 인공지능기술을 발판으로 하는 디지털 기반 혁신의 시대를 맞이하면서 다양한 영역에서 빅데이터가 발생, 수집, 축적, 공개되고 있으며, 수집된 대량의 데이터를 분석하여 데이터 항목들 간의 규칙성 및 관련성 등, 이용 가능한 지식을 추출하기 위한 데이터 분석, 데이터마이닝, 데이터사이언스 분야의 연구가 활발하게 진행되고 있다[4-6]. 특히, 다양한 데이터로부터 공통의 속성을 갖는 객체들을 그룹화하고 개념을 추출하고 계층구조를 구성하여 지식체계를 가시화하기 위한 형식개념분석기법[7]이 새로운 지식발견을 위한 다양한 분야에서 널리 응용되고 있다[8].

본 논문에서는, 형식개념분석기법을 확장하여, 객체의 속성뿐만 아니라 객체들 사이의 관련 관계를 기반으로 데이터집합을 분류하고 개념화하기 위한 기법을 제안한다. 구체적으로는, 주어진 데이터로부터 관계 속성을 기반으로 개념을 추출하여 계층 구조화하기 위한 관계형 개념분석기법의 제반 정의들을 토대로 개념분석알고리즘을 제안하고 분석도구(RCA Wizard)를 구현하여 몇 가지 공공데이터를 대상으로 수행한 분석실험 결과 등을 설명한다. 본 논문의 구성은 다음과 같다. 제2장에서는 형식개념분석기법에 대한 배경지식을 소개하고, 제3장에서는 관계 속성을 기반으로 하는 관계형 개념분석기법과 분석알고리즘 및 분석 도구를 제안한다. 제4장에서는 관계형 개념분석기법을 이용한 공공데이터 분석실험에 관해서 서술한다. 제5장에서는 본 논문의 결론과 향후 연구에 관해서 설명한다.

II. Related Works

형식개념분석기법[7]에서는 관심 영역 내의 분석 대상(객체. Object)들이 각각 어떠한 특징(속성. Attribute)을 소유하고 있는지를 테이블 형태로 정형화하여 나타낸 데이터 테이블(Formal Context)로부터 공통의 속성을 갖는 객체들을 개념(Concept) 단위로 그룹화하여 분류체계(Concept Hierarchy)를 구성해서 가시화하기 위한 수학적 이론체계를 제공한다. 형식개념분석기법의 수학적 이론체계를 구성하는 몇 가지 정의들은 다음과 같다.

[정의1]데이터테이블 $K=(G,M,I)$는 다음과 같은 3개 요소들로 구성되는 튜플이다.

- G : 객체들의 집합,
- M : 속성들의 집합,
- $I \subseteq G \times M$: 어떤 객체가 어떤 속성을 갖는지를 나타내는 관계 집합.

관계 집합의 임의의 원소$(g,m) \in I$는 객체g가 속성m을 가지고 있다는 것을 의미한다. Table 1은 5명의 학생과 6가지 속성들을 기반으로 어떤 학생이 어떠한 속성을 가졌는지를 나타낸 데이터 테이블$K_1=(G_1,M_1,I_1)$의 예이다.

Table 1. An example of Formal Context K_1

G_1 \ M_1	male	female	freshman	sophomore	junior	senior
s1	X		X			
s2	X		X			
s3		X			X	
s4		X		X		
s5	X					X

[정의2]데이터테이블 $K=(G,M,I)$가 주어졌을 때, 임의의 집합$X \subseteq G$와 $Y \subseteq M$에 대하여, 공통요소들을 추출하기 위한 연산자$CA: 2^G \rightarrow 2^M$와 $CO: 2^M \rightarrow 2^G$를 다음과 같이 정의한다.

- $CA(X)=\{m \in M \mid \forall g \in X: (g,m) \in I\}$,
- $CO(Y)=\{g \in G \mid \forall m \in Y: (g,m) \in I\}$.

연산자$CA(X)$는 집합X의 원소들(즉, 객체들)이 공유하고 있는 공통속성들의 집합을 추출하며, 연산자$CO(Y)$는 집합Y의 원소들(즉, 속성들)을 공유하고 있는 객체들의 집합을 추출해준다.

[정의3]데이터테이블 $K=(G,M,I)$의 임의의 두 집합 $X \subseteq G, Y \subseteq M$에 대하여,

$C=(X,Y) \Leftrightarrow (CA(X)=Y \wedge CO(Y)=X)$

를 만족하는 C를 데이터 테이블 $K=(G,M,I)$로부터 도출된 형식개념(또는 개념)이라고 한다.

즉, 데이터 테이블 $K=(G,M,I)$로부터 도출된 개념 C는, $CA(X)=Y \ \wedge CO(Y)=X$의 조건을 만족하는 두 집합$X \subseteq G, Y \subseteq M$로 구성되는 쌍$(X, Y)$이다. 이 때, X와 Y를 각각 개념C의 외연과 내포라고 부른다.

데이터 테이블 $K=(G,M,I)$로부터 도출된 모든 개념들의 집합을 $\mathbf{C}_K$라고 하고, $\mathbf{C}_K$의 임의의 두 개념 $C_1=(X_1,Y_1)$과 $C_2=(X_2,Y_2)$에 대하여, $X_1 \subseteq X_2 (\Leftrightarrow Y_2 \subseteq Y_1)$를 만족하는 경우, 두 개념들 사이에는 "상하위 개념관계"가 있다고 하고, $C_1 \leq_K C_2$로 나타낸다. 이때, C_1을 C_2의 하위개념, C_2를 C_1의 상위개념이라고 부른다. 데이터 테이블 $K=(G,M,I)$에 대한 형식개념들의 집합$\mathbf{C}_K$와 상하위 개념관계$\leq_K$는 완비격자(complete lattice)를 구성하며, $\mathbf{L}_K=(\mathbf{C}_K, \leq_K)$로 나타낸다. 이러한 완비격자를 개념 격자(Concept Lattice) 또는 개념 계층구조(Concept Hierarchy)라고 부른다. 데이터 테이블로부터 개념들을 추출하여 개념 계층구조를 구성하기 위한 알고리즘 FCA는 다음과 같다.

Algorithm FCA

입력: 데이터 테이블 $K=(G,M,I)$
출력: 개념 계층구조 $\mathbf{L}_K=(\mathbf{C}_K, \leq_K)$

1: **for all** $g \in G$ **do**
2: $\mathbf{C}_K \leftarrow \mathbf{C}_K \cup \{(CO(CA(g)), CA(g))\}$;
3: **for all** $c \in \mathbf{C}_K$ **do**
4: **for all** $g \in (G - CO(c))$ **do**
5: $X \leftarrow CO(c) \cup \{g\}$;
6: **if** $(CO(CA(X)), CA(X)) \notin \mathbf{C}_K$ **then**
7: $\mathbf{C}_K \leftarrow \mathbf{C}_K \cup \{(CO(CA(X)), CA(X))\}$;
8: **for all** $c_1 \in \mathbf{C}_K$ **do**
9: **for all** $c_2 \in (\mathbf{C}_K - \{c_1\})$ **do**
10: **if** $(c_1 \leq_K c_2) \wedge (\nexists c_3 \in \mathbf{C}_K - \{c_1, c_2\}[(c_1 \leq_K c_3) \wedge (c_3 \leq_K c_2)])$ **then**
11: $\leq_K \leftarrow \leq_K \cup \{(c_1, c_2)\}$;
12: **return** $\mathbf{L}_K=(\mathbf{C}_K, \leq_K)$;

개념 계층구조는 Fig. 1과 같은 형태의 개념계층구조도(Concept Hierarchy Diagram)로 가시화하며, 정점과 유향변은 각각 개념과 상하위개념관계를 나타낸다. 이때, 각 정점은 개념의 이름, 내포(해당 개념을 구성하는 객체들이 가지고 있는 속성들), 외연(해당 개념을 구성하는 객체들)을 구성요소로 하는 사각형으로 표시하고, 상하위개념관계는 하위개념으로부터 상위개념으로 향하는 화살표로 나타낸다. 특히, 개념 계층 구조도에서는, 속성들은 상위에서 하위로, 객체들은 하위에서 상위 방향으로 계승하는 형태로 간략화하여 표기한다. Table 1의 데이터 테이블 K_1에 형식개념분석기법을 적용하여 추출된 8개의 개념들은 상하위개념관계를 토대로 Fig. 1과 같은 개념계층구조도로 표현할 수 있다. 개념 계층구조도 상의 각 정점(개념)의 의미를 설명하기 위하여, 예를 들면, Fig. 1의 가장 왼쪽에 자리 잡고 있는 정점(Concept_Students_5)은, 2개의 속성(freshman, male)을 가지고 있는 2명의 학생(s1, s2)에 대응하는 개념(즉, 1학년 남학생)을 나타내고 있다. 또한, Concept_Students_1은, 2개의 속성(senior, male)을 가지고 있는 1개의 객체(s5)를 나타내는 개념(즉, 3학년 남학생)이다.

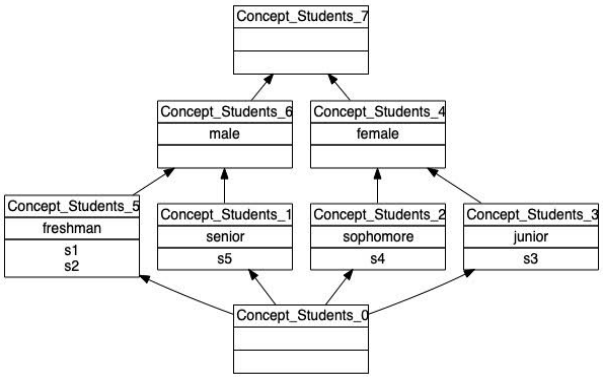


Fig. 1. Concept Hierarchy Diagram L_{K_1} for Table 1

형식개념분석에 의해 구축된 개념 계층 구조도는 주어진 데이터들에 대한 공통속성 기반의 분류체계를 제공한다. 예를 들면, Table 1의 데이터 테이블을 분석한 결과(Fig. 1)로부터, 6명의 학생은 남학생 그룹(Concept_Students_6 : s1, s2, s5)과 여학생 그룹(Concept_Students_4 : s3, s4)으로 분류되었다. 또한, 남학생그룹은 1학년 남학생 그룹(Concept_Students_5 : s1, s2)과 4학년 남학생 그룹(Concept_Students_1 : s5)으로, 여학생 그룹은 2학년 여학생 그룹(Concept_Students_2 : s4)과 3학년 여학생 그룹(Concept_Students_3 : s3)으로 다시 세부적으로 분류되었음을 알 수 있다.

III. The Proposed Method

형식개념분석기법은 관심 영역의 주어진 데이터셋으로부터 어떤 객체가 어떤 속성을 가졌는지를 나타내는 "객체-속성"관계(객체와 속성 사이의 소유관계)에 초점을 맞추어서 데이터를 분석하고 추출된 정보를 가시화한다. 그러나, 다양한 관심 영역들에 실존하는 대부분의 데이터에는 같은 영역 내에 존재하는 객체들 사이의 관련관계(예를 들

면, 학교 영역에서 교수-학생 관계) 뿐만 아니라, 서로 다른 영역에 소속된 객체들 사이의 관련 관계(예를 들면, 학생과 교과목 사이의 수강 관계)가 다수 존재하며, 데이터 분석과 데이터마이닝 분야에서는 이러한 관련 관계 데이터가 핵심 정보를 구성하고 새로운 지식을 발견하기 위한 중요한 자원이 된다[9].

위와 같은 형식개념분석기법의 한계점에 주목하여, 본 논문에서는 기존의 형식개념분석기법과 더불어서, 객체들 사이의 관련 관계 정보를 추출하여 분석하기 위한 관계형 개념분석기법(Relational Concept Analysis)[10]을 기반으로, 객체들 사이의 관련 관계를 토대로 데이터집합을 분류하고 개념화하기 위한 기법(개념분석 알고리즘 RCA와 분석 지원 도구 RCA Wizard)을 제안한다.

1. Definitions

관계형 개념분석기법의 제반 정의들[10]을 보다 구체적으로 정형화하여 아래의 정의4~6에 재정의하고, 이를 토대로 개념분석알고리즘(RScaling과 RCA)을 새롭게 제안하였다.

[정의4]관계형 데이터 테이블 그룹 $\Omega=(\boldsymbol{K},\boldsymbol{R})$는 데이터테이블들의 집합$\boldsymbol{K}$와 관련 관계 데이터 테이블들의 집합$\boldsymbol{R}$로 구성된 모임이다.

- $\boldsymbol{K}=\{K_i\}_{i\in[1,n]}$: $K_i=(G_i,M_i,I_i)$,
- $\boldsymbol{R}=\{R_j\}_{j\in[1,m]}$: $R_j=(G_k,G_l,r_j)$,

이때, $k,l\in[1,n]$에 대해서, $r_j\subseteq G_k\times G_l$은 관련 관계를 나타내며, r_j에 대한 정의역과 치역은 각각 $dom(r_j)=G_k$와 $ran(r_j)=G_l$이다.

관계형 데이터 테이블 그룹Ω은, 객체들의 속성 소유정보와 객체들 사이의 관련 관계 정보를, 각각 데이터 테이블들($\boldsymbol{K}$)과 관련 관계 데이터 테이블들($\boldsymbol{R}$)에 모아놓은 형태를 갖는다. 예를 들면, 관계형 데이터 테이블 그룹 $\Omega=(\{K_1,K_2\},\{R_1\})$은, Table 1의 학생 데이터 테이블$K_1=(G_1,M_1,I_1)$과 Table 2의 컴퓨팅기기에 관한 데이터 테이블$K_2=(G_2,M_2,I_2)$, 그리고 Table 3의 학생-컴퓨팅기기 소유관계($has\subseteq G_1\times G_2$)를 나타내는 관련 관계 데이터 테이블$R_1=(G_1,G_2,has)$로 구성된다.

관계형 개념분석기법에서는 관련 관계 데이터 테이블에 함축된 관련 관계 정보를 정의역에 해당하는 데이터테이블에 포함해서 표현하기 위하여 관계 정보 변환과정(relational scaling)을 거친다. 즉, 관계 정보 변환과정에서는 관련 관계$r_j\subseteq G_k\times G_l$를 정의역$dom(r_j)=G_k$의 데이터 테이블 내부에 다음과 같은 관계 속성(relational attribute)으로 추가하여 표현한다.

Table 2. An example of Formal Context K_2

G_2 \ M_2	Laptop	Desktop	Tablet	portable	Apple	Microsoft
MacBookAir	X			X	X	
iMac		X			X	
SurfacePro	X		X	X		X
iPad			X	X	X	

Table 3. An example of Relational Context R_1

has	MacBookAir	iMac	SurfacePro	iPad
s1	X		X	
s2	X			X
s3	X	X		X
s4		X	X	
s5		X		X

[정의5]관계형 데이터 테이블 그룹 $\Omega=(\boldsymbol{K},\boldsymbol{R})$의 임의의 관련 관계$r_j\subseteq G_k\times G_l$에 대하여, r_j의 정의역G_k의 객체($g\in G_k$)와 치역G_l의 객체들(치역의 데이터 테이블 K_l로부터 추출된 개념$c=(X,Y)\in\mathfrak{C}_{K_1}$의 외연을 형성하는 객체들) 사이의 관련 관계r_j가 $r_j(g)\cap X\neq\varnothing$의 조건을 만족하는 경우, 정의역의 데이터 테이블 $K_k=(G_k,M_k,I_k)$에 관계 속성$\exists r_j(c)$으로 정의한다.

관계형 데이터 테이블 그룹 $\Omega=(\boldsymbol{K},\boldsymbol{R})$의 임의의 관련 관계 데이터 테이블$R_j=(G_k,G_l,r_j)$과 이에 대응하는 관련 관계$r_j\subseteq G_k\times G_l$의 정의역 및 치역에 해당하는 데이터테이블$K_k=(G_k,M_k,I_k)$와 $K_l=(G_l,M_l,I_l)$에 대하여, 관계r_j를 기반으로 관계 속성$\exists r_j(c)$으로 K_k를 확장하기 위한 추가분(데이터 테이블$K_k^{r_j}$)은 다음과 같이 구성된다. 즉, $K_k^{r_j}=(G_k^{r_j},M_k^{r_j},I_k^{r_j})$,

- $G_k^{r_j}=G_k$,
- $M_k^{r_j}=\{\exists r_j(c)|c\in\mathfrak{C}_{K_1}\}$,
- $I_k^{r_j}=\{(g,\exists r_j(c))|g\in G_k,c=(X,Y)\in\mathfrak{C}_{k_i},r_j(g)\cap X\neq\varnothing\}$

앞서 주어진 데이터 테이블 그룹 $\Omega=(\{K_1,K_2\},\{R_1\})$에 대하여, 관계형 데이터 테이블 $R_1=(G_1,G_2,has)$을 기반으로 관계 속성을 추가하기 위한 데이터 테이블은 Table 4와 같다.

[정의6]주어진 데이터 테이블 $K_k=(G_k,M_k,I_k)$와 관련 관계$r_j\subseteq G_k\times G_l$를 기반으로 K_k를 확장하여 변환시킨 데이터 테이블 K_k^+을 다음과 같이 정의한다.

$K_k^+=(G_k^+,M_k^+,I_k^+)$,

- $G_k^+=G_k,\ M_k^+=M_k\cup M_k^{r_j},\ I_k^+=I_k\cup I_k^{r_j}$.

Table 4. Additional context $K_1^{has}=(G_1^{has}, M_1^{has}, I_1^{has})$ for weak relational scaling

G_1^{has} \ M_1^{has}	ㅠhas(C0)	ㅠhas(C1)	ㅠhas(C2)	ㅠhas(C3)	ㅠhas(C4)	ㅠhas(C5)	ㅠhas(C6)	ㅠhas(C7)	ㅠhas(C8)	ㅠhas(C9)	ㅠhas(C10)
s1			X		X	X	X	X	X	X	X
s2		X			X	X	X	X	X	X	X
s3		X		X	X	X	X	X	X	X	X
s4			X	X		X		X	X	X	X
s5		X		X		X	X		X	X	X

2. Algorithms

본 논문에서는 위의 정의들(정의4~6)을 토대로 관계 정보 변환과정(Relational Scaling)을 알고리즘**RScaling**으로 제안하였다. 알고리즘**RScaling**에서는, 입력으로 주어진 데이터 테이블$K_k=(G_k, M_k, I_k)$에 대하여 관련 관계 $r_j \subseteq G_k \times G_l$를 기반으로 K_k를 확장하여 변환시킨 $K_k^+=(G_k^+, M_k^+, I_k^+)$을 다음과 같은 절차에 의해 구성한다.

(1) G_k^+의 구성(라인1) :

입력으로 주어진 데이터 테이블$K_k=(G_k, M_k, I_k)$의 객체 집합G_k를 그대로 G_k^+로 유지한다.

(2) 치역의 데이터 테이블로부터 추출된 각 개념들 $c=(X, Y)\in \mathbf{C}_i$과 정의역의 객체$g\in G_k$에 대하여 아래의 처리를 반복적으로 수행한다.

(가) M_k^+의 구성(라인5) :

$K_k=(G_k, M_k, I_k)$의 속성집합M_k에 이항관계 $r_j \subseteq G_k \times G_l$를 기반으로 관계 속성$\exists r_j(c)$을 추가하여 M_k^+를 확장한다.

Algorithm RScaling	
입력	데이터 테이블$K_k=(G_k, M_k, I_k)$, 관련 관계 데이터 테이블$R_j=(G_k, G_l, r_j)$, 개념 계층구조들의 집합$L=\{(\mathbf{C}_i, \leq_i)\}_{i\in[1,n]}$
출력	관계 정보 변환 완료된 데이터 테이블 $K_k^+=(G_k^+, M_k^+, I_k^+)$
1:	$G_k^+ \leftarrow G_k$
2:	**foreach** $c=(X, Y)\in \mathbf{C}_i$ **and** $g\in G_k$ **do**
3:	**if** $(r_j(g)\cap X \neq \varnothing)$ **then**
4:	**foreach** $g\in G_k$ **and** $r_j \subseteq G_k \times G_l$ **do**
5:	$M_k^+ \leftarrow M_k \cup \{\exists r_j(c)\}$
6:	$I_k^+ = I_k \cup \{(g, \exists r_j(c))\}$
7:	**return** $K_k^+=(G_k^+, M_k^+, I_k^+)$

(가) I_k^+의 구성(라인6) :

$K_k=(G_k, M_k, I_k)$의 관계 집합I_k에 관련 관계 $r_j \subseteq G_k \times G_l$를 기반으로 $(g, \exists r_j(c))$을 추가하여 I_k^+를 확장한다.

데이터 테이블 그룹 $\Omega=(\{K_1, K_2\}, \{R_1\})$에 대하여, 위의 알고리즘을 적용한 결과(즉, **RScaling**(K_1, R_1, L_{K_2})을 수행한 결과)는 Table 5와 같다.

Table 5. Extended context $K_1^+=(G_1^+, M_1^+, I_1^+)$ based on relational attributes

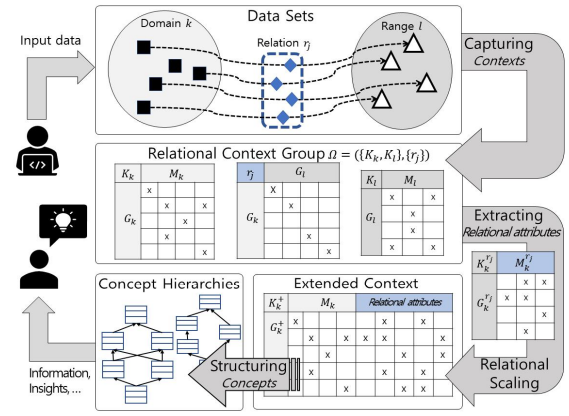
G_1^+ \ M_1^+	male	female	freshman	sophomore	junior	senior	ㅠhas(C0)	ㅠhas(C1)	ㅠhas(C2)	ㅠhas(C3)	ㅠhas(C4)	ㅠhas(C5)	ㅠhas(C6)	ㅠhas(C7)	ㅠhas(C8)	ㅠhas(C9)	ㅠhas(C10)
s1	X		X						X		X	X	X	X	X	X	X
s2	X		X					X			X	X	X	X	X	X	X
s3		X			X			X		X	X	X	X	X	X	X	X
s4		X		X					X	X		X		X	X	X	X
s5	X					X		X		X		X	X		X	X	X

이상에서 소개한 제반 정의들과 알고리즘 **FCA** 및 **RScaling**을 기반으로, 관계형 형식개념분석의 모든 과정(Fig. 2(a))을 알고리즘 **RCA**로 정리하여 제안한다.

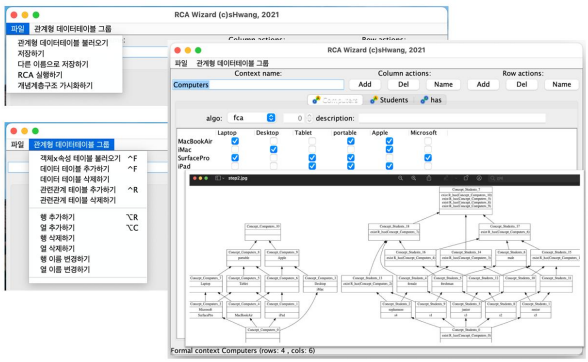
Algorithm RCA	
입력:	관계형 데이터 테이블 그룹 $\Omega=(\boldsymbol{K},\boldsymbol{R})$. (단, $\boldsymbol{K}=\{K_i\}_{i\in[1,n]}$, $\boldsymbol{R}=\{R_j\}_{j\in[1,m]}$)
출력:	개념 계층구조들의 집합L. (단, $L=\{L_i\}_{i\in[1,n]}=\{(\mathbf{C}_i, \leq_i)\}_{i\in[1,n]}$)
1:	$p \leftarrow 0$
2:	halt $\leftarrow$ ***false***
3:	**for** $i := 1..n$ **do**
4:	$L_i^0 \leftarrow$ FCA(K_i^0)
5:	**while not** halt **do**
6:	$p \leftarrow p+1$
7:	**for** $i := 1..n$ **do**
8:	$K_i^p \leftarrow$ **RScaling**(K_i^{p-1}, R_j, L^{p-1})
9:	$L_i^p \leftarrow$ **FCA**(K_i^p)
10:	**if** $\bigwedge_{i=1}^{n}(L_i^p = L_i^{p-1})$ **then** halt $\leftarrow true$

알고리즘 **RCA**에서는, 입력으로 주어진 관계형 데이터 테이블 그룹$\Omega=(\boldsymbol{K},\boldsymbol{R})$의 데이터 테이블들의 집합 $\boldsymbol{K}=\{K_i\}_{i\in[1,n]}$(단, $K_i=(G_i, M_i, I_i)$)의 각 데이터 테

이블에 대해서 형식개념분석기법을 적용하여 개념 계층구조를 구성한 후에(3~4라인), (5~10라인)의 과정을 반복하면서, 전 단계($p-1$)에서 구성된 개념으로부터 도출된 관계 속성을 기반으로 데이터 테이블을 확장시키는 관계 정보 변환과정을 수행한 후(8라인)에 형식개념분석기법을 적용하여 새로운 개념 계층구조를 구성하는 작업(9라인)을 시행하면서 전후 단계($p-1, p$단계)에서 구해지는 개념 계층구조들이 동형(isomorphic), 즉, 데이터 테이블에 변화가 없는 동일한 상태가 되면 종료한다(10라인).



(a)



(b)

Fig. 2. Overview of Processing of RCA(a) and screenshots of RCA Wizard(b)

3. RCA Wizard

본 논문에서 제안한 알고리즘 **RCA**를 토대로 관계형 형식개념분석 지원도구 **RCA Wizard**를 개발하였다. RCA Wizard는 Eclipse IDE 2021-03 개발환경과 Java(14.0.2 버전) 기반으로 제작되었으며, 아래와 같은 기능들(메뉴들)이 Fig. 2(b)와 같은 인터페이스 형태로 구성되어 있다.

- 파일
  - 관계형 데이터 테이블 불러오기
  - 저장하기
  - 다른 이름으로 저장하기
  - RCA 실행하기
  - 개념 계층구조 가시화하기
- 관계형 데이터 테이블 그룹
  - 객체x속성 테이블 불러오기
  - 데이터 테이블 추가/삭제
  - 관련관계 테이블 추가/삭제
  - 행/열의 추가/삭제
  - 행/열의 이름 변경

사용자는 “관계형 데이터 테이블 그룹” 메뉴의 각 서브메뉴들을 사용하여 관계형 데이터 테이블 그룹을 입력(또는 편집)하고, “파일”메뉴의 “RCA 실행하기” 서브메뉴를 선택하면 본 논문에서 제안한 RCA 알고리즘이 실행되어 개념 계층구조들이 생성된다. 이어서 “개념 계층구조 가시화하기” 서브메뉴를 실행하면 개념 계층구조가 가시화되어 화면에 표시된다.

Fig. 3은 앞서 살펴보았던 관계형 데이터테이블 그룹 $\Omega = (\{K_1, K_2\}, \{R_1\})$에 대하여 **RCA Wizard**를 이용하여 **RCA**(Ω)을 실행한 결과물(개념 계층구조들)이다. Fig. 3의 최종결과물(개념 계층구조도 $L^2_{K_1^+}$)의 주요 정점들은 다음과 같은 정보를 나타내고 있다.

- **Concept_Students_13** : Microsoft의 SurfacePro와 같은 랩탑과 태블릿 겸용 제품을 가지고 있는 학생들은 s1과 s4이다.
- **Concept_Students_4** : s4와 s3와 같은 여학생들을 나타내는 정점(개념)으로써, 이 학생들은 SurfacePro와 MacBookAir와 같은 랩탑과 iMac과 같은 Apple의 데스크톱을 가지고 있다.
- **Concept_Students_5** : s1과 s2와 같은 남자 신입생들을 나타내는 정점(개념)으로서, Apple의 휴대용 랩톱 MacBookAir를 소유하고 있다는 공통점이 있다.
- **Concept_Students_12** : iPad와 MacBookAir를 모두 가지고 있는 학생들은 s3와 s2이다.
- **Concept_Students_10** : 학생 s3와 s5는 Apple의 휴대용 iPad와 데스크톱 iMac을 소유하고 있다는 공통점이 있다.
- **Concept_Students_11** : s2와 s5는 남학생들로서, Apple의 휴대용 iPad를 소유하고 있다는 공통점이 있다.
- **Concept_Students_16** : MacBookAir와 같은 Apple의 휴대용 랩탑을 가지고 있는 학생들은 s1, s2, s3이다.
- **Concept_Students_14** : iMac과 같은 Apple의 데스크톱을 가지고 있는 학생들은 s3, s4, s5이다.

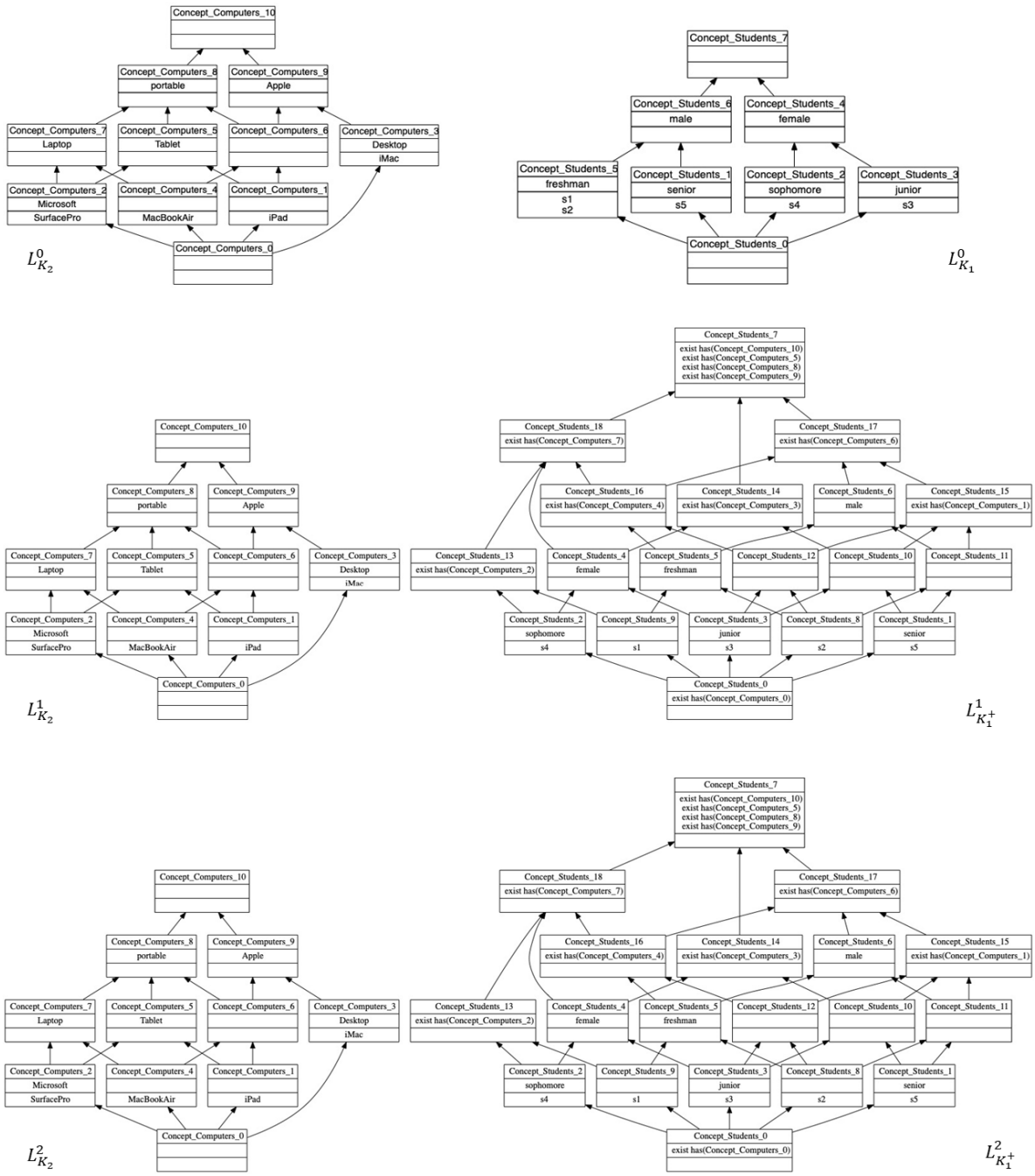
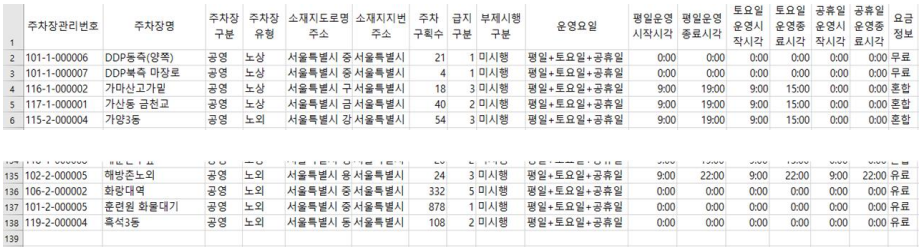


Fig. 3. Results of **RCA(Ω)**, where $\Omega = (\{K_1, K_2\}, \{R_1\})$

- **Concept_Students_6** : 휴대용 Apple기기를 가지고 있는 남학생그룹을 나타내는 정점으로써, 이 그룹에는 s1, s2, s5와 같은 학생들이 포함되어 있다.
- **Concept_Students_15** : iPad와 같은 Apple의 휴대용 태블릿을 가지고 있는 학생 그룹을 나타내는 정점으로써, 이 그룹에는 s3와 s5가 포함되어 있다.
- **Concept_Students_18** : 휴대용 랩톱을 가지고 있는 학생 그룹을 나타내는 정점으로써, 이 그룹에는 s1, s2, s3, s4와 같은 학생들이 포함되어 있다.
- **Concept_Students_17** : 휴대용 Apple 제품을 소유한 학생 그룹을 나타내는 정점으로써, 이 학생 그룹은 다시 MacBookAir를 소유한 학생 그룹(Concept_Students_16 : s1, s2, s3)과 iPad를 소유한 학생 그룹(Concept_Students_15 : s2, s3, s5), 그리고 남학생그룹(Concept_Students_6 : s1, s2, s5)으로 분류된다.



	주차장관리번호	주차장명	주차장 구분	주차장 유형	소재지도로명 주소	소재지지번 주소	주차 구획수	급지 구분	부제시행 구분	운영요일	평일운영 시작시각	평일운영 종료시각	토요일 운영시 작시각	토요일 운영종 료시각	공휴일 운영시 작시각	공휴일 운영종 료시각	요금 정보
2	101-1-000006	DDP동측(양쪽)	공영	노상	서울특별시 중	서울특별시	21	1	미시행	평일+토요일+공휴일	0:00	0:00	0:00	0:00	0:00	0:00	무료
3	101-1-000007	DDP북측 마장로	공영	노상	서울특별시 중	서울특별시	4	1	미시행	평일+토요일+공휴일	0:00	0:00	0:00	0:00	0:00	0:00	무료
4	116-1-000002	가마산고가밑	공영	노상	서울특별시 구	서울특별시	18	3	미시행	평일+토요일+공휴일	9:00	19:00	9:00	15:00	0:00	0:00	혼합
5	117-1-000001	가산동 금천교	공영	노상	서울특별시 금	서울특별시	40	2	미시행	평일+토요일+공휴일	9:00	19:00	9:00	15:00	0:00	0:00	혼합
6	115-2-000004	가양3동	공영	노외	서울특별시 강	서울특별시	54	3	미시행	평일+토요일+공휴일	9:00	19:00	9:00	15:00	0:00	0:00	혼합
135	102-2-000005	해방촌노외	공영	노외	서울특별시 용	서울특별시	24	3	미시행	평일+토요일+공휴일	9:00	22:00	9:00	22:00	9:00	22:00	유료
136	106-2-000002	화랑대역	공영	노외	서울특별시 중	서울특별시	332	5	미시행	평일+토요일+공휴일	0:00	0:00	0:00	0:00	0:00	0:00	유료
137	101-2-000005	훈련원 화물대기	공영	노외	서울특별시 중	서울특별시	878	1	미시행	평일+토요일+공휴일	0:00	0:00	0:00	0:00	0:00	0:00	유료
138	119-2-000004	흑석3동	공영	노외	서울특별시 동	서울특별시	108	2	미시행	평일+토요일+공휴일	0:00	0:00	0:00	0:00	0:00	0:00	유료
139																	

Fig. 4. Public Open Data of public parking lot of Seoul Facilities Corporation

IV. Experiments and Results

우리나라의 공공데이터포탈(www.data.go.kr)에서는 교육, 국토관리, 공공행정 등 총 16개 영역으로 분류하여 956개 기관에서 38,139개의 파일데이터와 6,774개의 오픈 API를 개방하고 있다. 본 연구에서는 관계형 형식개념분석기법(RCA 알고리즘과 분석 지원 도구 **RCA Wizard**)의 유효성을 실증하기 위하여, 이러한 공공데이터 중에서 조회 수 2만회를 넘고 있는 전국주차장정보표준데이터(2021.4.1.현재)[2]를 대상으로 **RCA Wizard**를 이용하여 다음과 같은 실험들(Experiment 1, Experiment 2)을 실시하였다.

1. Experiment 1

1.1 Dataset

전국주차장정보표준데이터 중에서 “서울시설공단”에서 관리하는 137개소의 공영주차장 관련 데이터(Fig. 4)를 추출하여 주차장 유형(노상/노외)과 요금정보(유료/무료/혼합)와 같은 중요 속성들을 기반으로 Table 6과 같은 서울시설공단 공영주차장 데이터 테이블K_1을 구성하였다. 또한, 서울시 생활권 계획[3]에서 공개하고 있는 권역생활권 구분자료(Table 7)를 토대로 25개의 객체와 5개의 속성을 갖는 서울시 권역생활권 데이터 테이블K_2(Table 8)를 구성하였다. 한편, 각 주차장의 소재지 지번 주소를 바탕으로, K_1의 각 주차장이 서울의 어떤 행정자치구에 자리 잡고 있는지를 나타내는 관련 관계($At \subseteq G_1 \times G_2$)를 기반으로 관련관계 데이터테이블$R_{At}$을 구성하였다(Table 9).

Table 6. Formal context K_1 of Fig. 5

G_1 \ M_1	노상	노외	유료	무료	혼합
DDP동측(양쪽)	X			X	
DDP북측마장로	X			X	
가마산고가밑	X				X
가산동 금천교	X				X
가양3동		X			X
⋮	⋮	⋮	⋮	⋮	⋮
해방촌노외		X	X		
화랑대역		X	X		
훈련원 화물대기		X	X		
흑석3동		X	X		

Table 7. Division of living areas of Seoul Metropolitan Government

권역구분	자치구
도심권	종로구, 중구, 용산구
동북권	성동구, 광진구, 동대문구, 중랑구, 성북구, 강북구, 도봉구, 노원구
서북권	은평구, 서대문구, 마포구
서남권	양천구, 강서구, 구로구, 금천구, 영등포구, 동작구, 관악구
동남권	서초구, 강남구, 송파구, 강동구

Table 8. Formal context K_2 of Table 7

G_2 \ M_2	도심권	동북권	서북권	서남권	동남권
종로구	X				
중구	X				
용산구	X				
성동구		X			
광진구		X			
⋮	⋮	⋮	⋮	⋮	⋮
강남구					X
송파구					X
강동구					X

2) https://www.data.go.kr/data/15012896/standard.do

3) https://planning.seoul.go.kr/plan/main.do

Table 9. Relational context R_{At} of relationships between parking lots and their locations

G_1 \ G_2	종로구	중구	용산구	...	중랑구	...	강서구	구로구	금천구	...	동작구	...
DDP동측(양쪽)		X		...		...				...		...
DDP북측 마장로		X		...		...				...		...
가마산 고가밑				...		...		X		...		...
가산동 금천교				...		...			X	...		...
가양3동				...		...	X			...		...
⋮	⋮	⋮	⋮	...	⋮	...	⋮	⋮	⋮	...	⋮	⋮
해방촌 노외			X	...		...				...		...
화랑대역				...	X	...				...		...
훈련원 화물대기		X		...		...				...		...
흑석3동				...		...				...	X	...

1.2 Experimental Process

위와 같은 데이터 테이블K_1, K_2, R_{At}에 대하여, K_1의 객체(서울시 공영주차장) 개수를 10개씩 증가시킨 단위로 총 14개의 데이터 테이블($K_1^1, K_1^2, ..., K_1^{14}$)을 구성하고, 각각에 대하여 **RCA Wizard**를 이용하여 관계형 데이터테이블 그룹$\Omega^i = (\{K_1^i, K_2\}, \{R_{At}\})$(단, $i = 1 ... 14$)에 대한 관계형 형식개념분석 알고리즘 **RCA**(Ω^i)을 적용하는 실험을 총 14회 실시하였다.

1.3 Results

Fig. 5는 14회의 실험 결과물 중에서 일부분(실험1-1, 실험1-2, 실험1-3, 실험1-4)을 나타내고 있다[4]. 각각의 개념 계층 구조도는 다양한 정보를 함의하고 있다. 예를 들어, Fig. 5의 실험1-1에 대한 실행 결과(L_{K_1} of Experiment 1-1)에서 가장 왼쪽에 있는 정점(Concept_공영주차장_4)은, 10개의 공영주차장 중에서 도심권(종로구, 중구, 용산구)에 있는 노상무료 공영주차장은 DDP동측(양쪽)과 DDP북측 마장로와 같은 2개소가 존재하고 있음을 나타내고 있다. 또한, 정점(Concept_공영주차장_11)은, 10개의 공영주차장 중에서 서남권에 있는 노상 공영주차장을 나타내는 개념으로써, 개봉역(북), 가마산고가밑, 가산동 금천교, 개봉역(남단), 개봉역(중앙)과 같은 주차장들이 이에 해당한다는 사실을 나타내고 있다.

이어서, 20개의 공영주차장을 대상으로 실험을 수행한 결과(L_{K_1} of Experiment 1-2)에서는, 더욱 풍부한 정보를 갖는 개념들을 더 많이 발굴할 수 있다. 예를 들면, L_{K_1} of Experiment 1-2의 정점(Concept_공영주차장_4)은 도심권(종로구, 중구, 용산구)에 있는 노상무료 공영주차장을 나타내고 있으며, 정점(Concept_공영주차장_13)은 도심권(종로구, 중구, 용산구)에 있는 노상 유료 공영주차장을 나타내고 있다. 이처럼 실험 대상이 되는 데이터테이블의 객체 개수가 증가할수록 실험 결과는 더욱 다양하고 풍부한 정보를 함의하게 되며, 더욱 세부적인 데이터분류 및 계층화가 가능하다.

또한, 실험1-1~14의 결과로부터, 실험데이터(즉, K_1의 객체 집합 크기 : $|G_1|$)가 증가함에 따라서, 관련 관계 $At \subseteq G_1 \times G_2$의 정의역에 해당하는 개념 계층구조$L_{K_1}$에서 발굴되는 개념들의 규모는 일정한 수에 수렴하고, 치역에 해당하는 개념 계층구조L_{K_2}의 개념들의 규모는 변화 없음을 알 수 있다(Fig. 6).

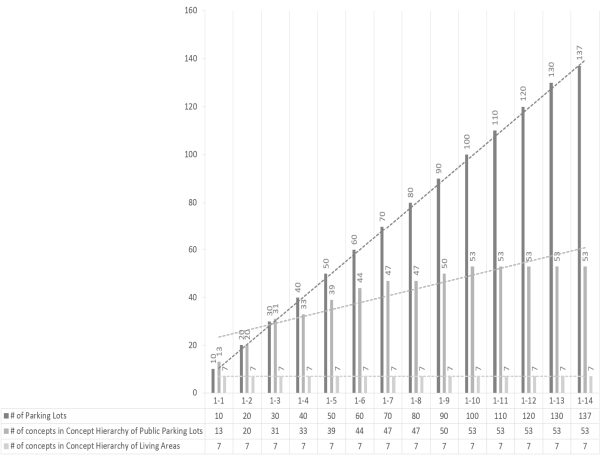


Fig. 6. Number of concepts in Experiment 1-1 to 1-14

2. Experiment 2

1.1 Dataset

두 번째 실험에서는, 서울시설공단에서 관리하는 137개 공영주차장 데이터에 대하여, 각 주차장의 소재지 지번 주소를 기반으로 서울시의 25개 행정자치구별로 분류하고, 각 주차장의 유형(노상/노외)과 요금정보(유료/무료/혼합)와 같은 속성들을 기반으로 총 25개의 행정자치구별 공영주차장 데이터 테이블을 구성하였다. Table 10은 서울시설공단에서 관리하는 강남구에 있는 9개 공영주차장 데이터 테이블K_1의 예를 나타내고 있다.

또한, 서울시의 도시기본계획("2030 서울생활권계획")

4) 지면 관계상, 총 14회의 모든 실험 결과물은 아래의 링크에 게시하였음.
https://drive.google.com/drive/folders/1TwnQ260p2jvxmdE2IFSznvDSdLx38QI6?usp=sharing

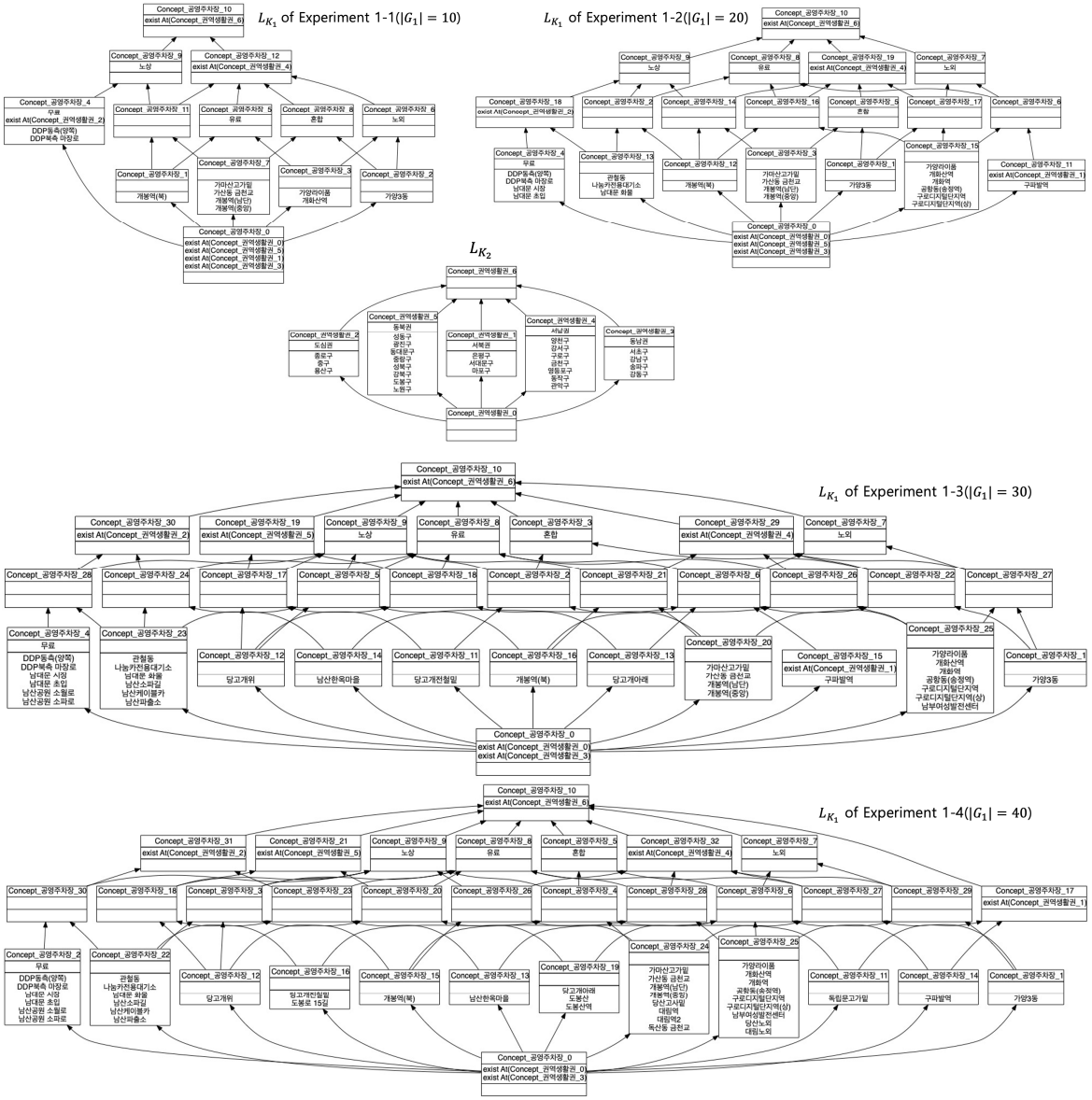


Fig. 5. Concept Hierarchy Diagrams of Experiment 1-1 ~ 1-4

Table 10. Formal context K_1 of Public Parking Lots of Gangnam-Gu

G_1 \ M_1	노상	노외	유료	무료	혼합
동호대교(남)	X				X
수서역		X	X		
압구정고가1	X		X		
압구정고가2	X		X		
영동6교밑	X		X		
일원동		X			X
일원역		X			X
일원터널		X	X		
학여울역		X	X		

에서는, 서울시 전역을 5개 권역생활권(경제활동이 이루어지는 대권역)과 116개 지역생활권(통근 통학 쇼핑 여가 등 주민들의 일상적인 생활 활동이 이루어지는 소권역)으로 구분하고, 관련 상세데이터를 서울열린데이터광장(https://data.seoul.go.kr/index.do)에서 “서울시 생활권 계획 권역 구분정보”(http://data.seoul.go.kr/dataList/OA-12074/S/1/datasetView.do)로 공개하고 있다. 본 실험(Experiment 2)에서는 서울시 25개 각 행정자치구에 대하여, 해당 구의 지역생활권을 구성하는 행정동에 관한 데이터를 추출하여 지역생활권 데이터 테이블을 구성하였다. Table 11은 강남구의 6개 지역생활권이 어떤 행정동으로 구성되어 있는지를 나타내는 데이터테이블K_2의 예이다.

한편, 서울시설공단에서 관리하는 서울시 25개 행정자치구별 각 공영주차장의 소재지 지번 주소를 기반으로, 각 행정자치구별 주차장들(G_1)이 서울의 어떤 지역생활권의

Table 11. Formal context K_2 of Local Living Areas of Gangnam-Gu

G_2 \ M_2	개포·일원	대치·도곡	삼성	수서·세곡	압구정·청담	역삼·논현
개포1동	X					
개포2동	X					
개포4동	X					
일원1동	X					
일원2동	X					
대치1동		X				
대치2동		X				
대치4동		X				
도곡1동		X				
도곡2동		X				
삼성1동			X			
삼성2동			X			
세곡동				X		
수서동				X		
일원본동				X		
신사동					X	
압구정동					X	
청담동					X	
논현1동						X
논현2동						X
역삼1동						X
역삼2동						X

Table 12. Relational context R_{At} of relationships between parking lots and their locations in local living areas of Gangnam-Gu

G_1 \ G_2	개포1동	개포2동	개포4동	일원1동	일원2동	대치1동	대치2동	…	수서동	일원본동	신사동	압구정동	…
동호대교(남)												X	
수서역									X				
압구정고가1											X		
압구정고가2											X		
영동6교밑							X	…					…
일원동				X									
일원역										X			
일원터널										X			
학여울역							X						

동(G_2)에 위치해 있는지를 관련 관계($At \subseteq G_1 \times G_2$)로 파악하여, 25개의 관련 관계 데이터 테이블을 구성하였다. Table 12는 강남구에 있는 서울시설공단 공영주차장이 강남구의 어떤 지역생활권에 위치하고 있는지를 나타내고 있는 관련 관계 데이터 테이블R_{At}의 예이다.

1.2 Experimental Process

본 실험에서는 25개 각 행정자치구의 데이터 테이블 K_1, K_2, R_{At}에 대하여 관계형 데이터 테이블 그룹 $\Omega = (\{K_1, K_2\}, \{R_{At}\})$을 구성하고 **RCA Wizard**를 이용하여 총 25회의 관계형 형식개념분석 실험을 실시하였다. 즉, 25회의 실험 중에서, Table 10, 11, 12를 기반으로 하는 강남구에 관한 관계형 데이터 테이블 그룹 $\Omega_{강남구} = (\{K_1, K_2\}, \{R_{At}\})$에 대한 실험 결과(즉, **RCA($\Omega_{강남구}$)**을 시행한 결과로부터 구축된 개념 계층구조도)는 Fig. 7과 같다[5].

1.3 Results

실험 결과물(Fig. 7의 개념 계층구조도L_{K_1})로부터 아래와 같은 다양한 정보를 파악할 수 있다.

- **Concept_강남구공영주차장_15** : 강남구지역생활권_2(압구정·청담)의 노상 유료주차장을 나타내는 개념으로써, 압구정고가1주차장과 압구정고가2 주차장이 있다.
- **Concept_강남구공영주차장_6** : 강남구에 있는 노상주차장을 나타내는 개념으로써, 총 4개소(압구정고가1, 압구정고가2, 동호대교(남), 영동6교밑)가 존재하며, 2개소(압구정고가1, 압구정고가2)는 강남구지역생활권_2(압구정·청담)에 있고 유료주차장이며, 1개소(동호대교(남))는 강남구지역생활권_2(압구정·청담)에 위치하고 있고 요금체계는 혼합형이다. 또한, 1개소(영동6교밑)는 강남구지역생활권_5(대치·도곡)에 위치하고 있는 노상 유료주차장이다.
- **Concept_강남구공영주차장_3** : 강남구에 있는 노상 유료주차장을 나타내고 있는 개념으로서, 3개의 주차장(압구정고가1, 압구정고가2, 영동6교밑)이 있다.

또한, Fig. 8은, 서울시설공단에서 관리하는 137개 공영주차장을 25개 행정자치구별로 분류하여, 각 행정자치구에 대한 공영주차장 수, 지역생활권 수, 그리고 본 실험에서 수행된 알고리즘 RCA의 내부반복 횟수(즉, 라인 5~10의 반복횟수), 그리고, 최종적으로 발견된 개념 수 등을 종합한 결과를 나타내고 있다. 예를 들어, Fig. 9에서 강남구의 경우, 서울시설공단에서 관리하는 공영주차장은 총 9개소이고, 강남구 내에 지역생활권은 총 6개 구역으로 구성되어 있으며, 이번 실험에서는 알고리즘 RCA의 라인 5~10이 총 3회 반복되어 최종적으로 19개의 개념이 발굴되었다. 또한, Fig. 9로부터, 강서구, 강남구, 중구, 구로구, 노원구 등의 순으로 각각 20개, 19개, 18개, 17개, 16개의 개념이 발견되었으며 평균 2.68회의 내부 반복처리에 의해 총 208개의 개념을 추출하였음을 알 수 있다.

5) 지면 관계상, 총 25회의 모든 실험 결과는 아래의 링크에 게시하였음.
https://drive.google.com/drive/folders/12I5N5oLN-ZGfJdMIYO5E5nzib56hq5Wv?usp=sharing

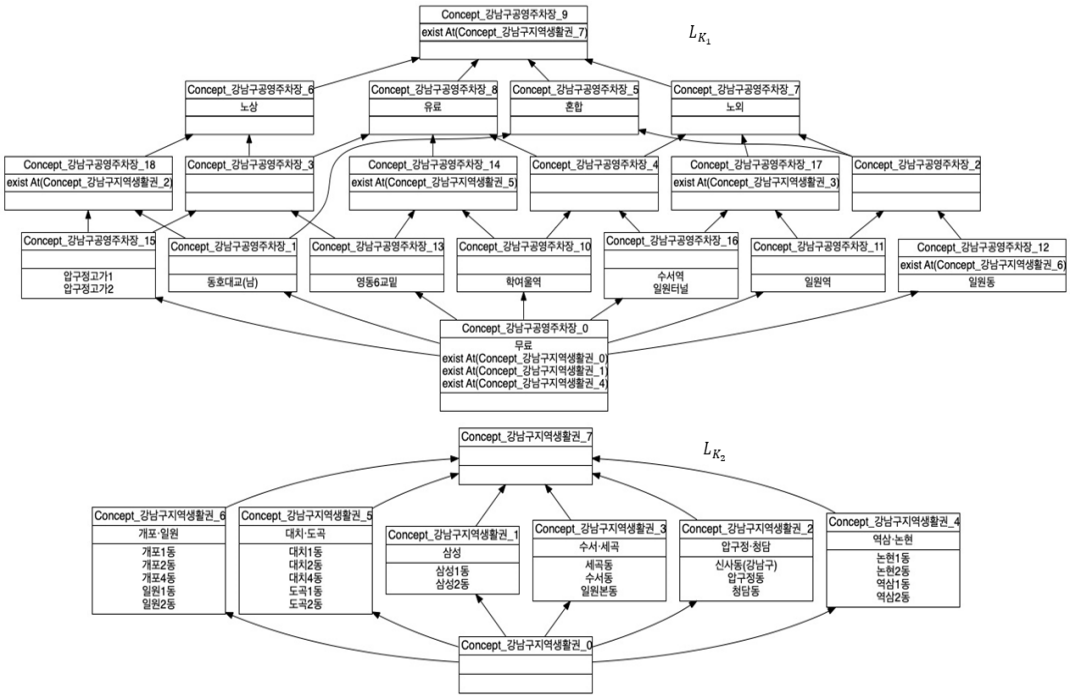


Fig. 7. Concept Hierarchy Diagrams of Experiment 2

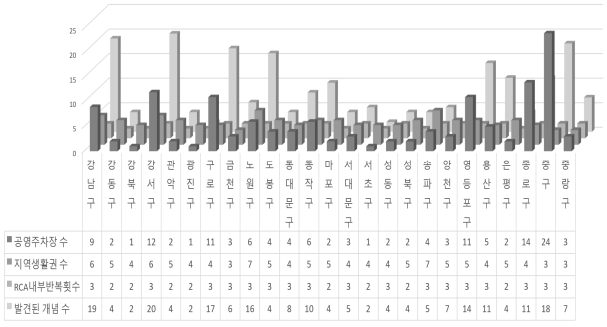


Fig. 8. Number of Public Parking Lots, Local Living Areas, Iterations and Concepts from Experiment 2

V. Conclusions

데이터 중심의 4차 산업혁명 시대를 맞이하여, 다종다양한 데이터에 내재되어 있는 정보를 추출하고 분석하여 그 결과를 가시화하기 위한 데이터분석기법에 대한 관심이 높아지면서 최근에는 형식개념분석기법이 주목받고 있다. 형식개념분석기법에서는, 주어진 데이터로부터 객체들의 공통속성을 기반으로 개념을 추출하여 계층화된 분류체계를 가시화할 수 있으나, 객체들 사이의 관련 관계를 분석할 수 없는 한계점이 존재한다. 본 논문에서는, 객체들의 공통속성뿐만 아니라 객체들 사이의 관련 관계를 바탕으로 개념을 추출하여 계층 구조화하기 위하여, 관계형 개념분석기법의 제반 정의들을 기반으로 개념분석 알고리즘을 제안하고 분석 지원 도구 RCA Wizard를 개발하였다. 또한, 공공데이터포털에서 개방 중인 몇 가지 공공데이터(서울시설공단 공영주차장데이터, 서울시 생활권 계획권역구분정보)를 대상으로 분석실험을 수행하여 연구 결과의 유용성을 실증하였다.

본 논문의 연구 결과는 기존의 관련 연구[10]와 대비하여 다음과 같은 차별성을 갖는다.

- 기존의 관련 연구[10]에서는 관계형 개념분석기법의 기본적인 정의들이 제안되었으나, 본 논문에서는 보다 구체적으로 제반 정의들을 정형화하고 이를 토대로 관계형 개념분석 알고리즘을 새롭게 제안하여, 분석지원도구 RCA Wizard를 개발하였다.
- 본 연구에서는 공공 오픈 데이터를 대상으로 데이터의 수집, 분석실험의 실시, 데이터의 분류와 군집화 및 내재되어 있는 정보의 시각화 등과 같은 일련의 데이터분석 및 가시화에 관한 실험을 수행하여 제안한 알고리즘과 분석 도구의 유용성을 실증하였다.

본 논문에서는 분석 결과(개념 계층구조)의 가시화에 초점을 맞추고 있으나, 더욱 다양하고 풍부한 분석 결과를 도출하기 위해서는 개념 계층구조로부터 연관규칙을 추출과 추론시스템을 개발하기 위한 후속 연구가 반드시 필요하다. 한편, 본 연구에서는 공공데이터포털에서 제공하는 데이터를 활용하고 있으나, 현존하는 국내의 공공데이터포털(www.data.go.kr)에서는 공공데이터의 개방, 국가데이터맵의 제공, 공공데이터의 제공요청, 데이터활용(시각화,

국민참여지도, 위치정보시각화) 등을 중심으로 운영되고 있어서 다양한 방법에 의한 데이터분석과 가시화가 미흡하며 공공포털데이터 내에 포함된 데이터의 다양성과 복잡성을 반영할 수 있는 분석 및 가시화 기법이 필요하다. 따라서, 본 연구결과와 후속연구들을 기반으로 향후에 공공데이터 포탈에서 공공데이터를 대상으로 하는 보다 효과적인 데이터분석, 분류, 군집화, 시각화, 정보검색 등에 활용할 수 있을 것으로 예상한다.

ACKNOWLEDGEMENT

This paper is the result of works performed during the 2021 research year of SunMoon University.

REFERENCES

[1] Korea Data Agency, "2020 Data Industry White Paper," Korea Data Agency, Vol. 23, pp. 4-5, 2020.
[2] Public Data Portal, https://www.data.go.kr
[3] H.D. Moon, "2021 Digital Innovation Outlook by DNA(Digital, Network, AI)," Monthly Software Oriented Society, No.79, pp.25-26, 2021.
[4] Gupta, M.K., Chandra, P., "A comprehensive survey of data mining," International Journal of Information Technology, 12, pp.1243-1257, 2020. DOI: 10.1007/s41870-020-00427-7
[5] L. Cao, "Data Science: A Comprehensive Overview," ACM Computing Surveys, 50(3), pp.1-42, 2017. DOI: 10.1145/3076253
[6] M. U. Raza and Z. XuJian, "A Comprehensive Overview of BIG DATA Technologies: A Survey," Proceedings of the 2020 5th International Conference on Big Data and Computing, New York, NY, USA, pp.23-31, 2020. DOI: 10.1145/3404687.3404694
[7] B. Ganter, and R. Wille, "Formal Concept Analysis: Mathematical Foundations," Springer, pp.1-15, 1999.
[8] P. K. Singh, C. Aswani Kumar, A. Gani, "A comprehensive survey on formal concept analysis, its research trends and applications," International Journal of Applied Mathematics and Computer Science, 26(2), pp. 495-516, Jun. 2016. DOI: 10.1515/amcs-2016-0035
[9] Džeroski S. "Relational Data Mining," Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA. 2009. DOI: 10.1007/978-0-387-09823-4_46
[10] Hacene, M.R., M. Huchard, A. Napoli, P. Valtchev, "Relational concept analysis: mining concept lattices from multi-relational data," Annals of Mathematics and Artificial Intelligence, 67(1): pp.81-108, 2013. DOI: 10.1007/s10472-012-9329-3

Authors

Suk-Hyung Hwang received the B.S degree in Computer Science from Kangwon National University in 1991, the M.E. and Ph.D. degrees in Information and Computer Science from Osaka University, Japan, in 1994 and 1997, respectively. Dr. Hwang joined the faculty of the Department of Computer Science and Engineering at SunMoon University, Korea, in 1997. He is currently a Professor in the Department of Artificial Intelligence and Software Technology, SunMoon University. His research interests include Object-Oriented Software Engineering, Data Science, Formal Concept Analysis, etc.