입자화 정도를 기반으로 하는 개념계층구조의 구축

강유경*, 황석형*
*선문대학교 컴퓨터공학과
e-mail:{aquamint99, shwang}@sunmoon.ac.kr

A Study on Construction of Granular Concept Hierarchies based Granularity Level

Yu-Kyung Kang*, Suk-Hyung Hwang*
*Dept of Computer Science & Engineering, SunMoon University

요 약

형식개념분석기법(FCA : Formal Concept Analysis)은 주어진 데이터로부터 공통속성을 갖는 객체들을 클러스터링하여 정보의 최소단위로써 개념(Concept)들을 추출하고 그들 사이의 관계를 토대로 계층화하여 데이터에 내재된 개념들의 구조를 가시화 해주는 Granular Computing의 한 종류이다. 형식개념분석기법에서는 공통속성을 갖는 객체들을 추출한다는 전제조건을 토대로 개념을 추출하기 때문에 다양한 상황이나 조건에 적합한 새로운 개념들을 추출하기에는 한계가 있다. 이와 같은 문제를 해결하기 위한 한 가지 방법으로써, 본 논문에서는 입자화 정도(granularity level)를 기반으로 하는 형식개념분석기법을 제안한다. 본 논문에서 제안하는 기법에서는 형식개념분석기법에 입자화 정도를 도입하여 다양한 조건과 추상화 수준을 토대로 하여, 개념들을 추출하고 개념계층구조를 구축할 수 있다.

1. 서론

Granular Computing은 주어진 문제를 해결하기 위해 입자(granule)들을 사용하는 모든 분야에 대한 포괄적인 용어이다. 하나의 입자는 주어진 도메인에 포함된 객체들의 부분집합 중의 하나를 의미하며, 주어진 도메인 안에 모든 객체들을 입자들로 분류하는 것을 입자화(granulation)라 부른다. 주어진 도메인에 대한 입자화는 객체들을 클러스터들로 그룹핑하거나 도메인의 객체들을 부분집합들로 분할하는 것으로, 데이터를 분석하고자하는 관점에 따라 다양한 입자들이 생성될 수 있다. Granular Computing을 기반으로 하는 데이터 분석 분야에는 형식개념분석기법(FCA : Formal Concept Analysis), Rough Set Analysis(RSA), Fuzzy Set Analysis(FSA) 등이 있다[1-5].

형식개념분석기법은 주어진 데이터로부터 공통속성을 갖는 객체들을 클러스터링하여 정보의 최소단위로써 개념(Concept)들을 추출하고 그들 사이의 관계를 토대로 계층화하여 데이터에 내재된 개념들의 구조를 가시화 해주는 Granular Computing의 한 종류이다. 형식개념분석기법의 결과물로는 개념들과 그 순서관계에 의해 구성되는 개념계층구조(Concept Lattice)와 개념계층구조를 토대로 속성들을 기반으로 하는 객체들 사이의 연관규칙(Association rules) 등을 추출할 수 있기 때문에 데이터 분석을 위한 많은 분야에서 유용하게 사용되고 있다[3]. 형식개념분석기법을 기반으로 다양한 데이터를 분석하기 위한 기법들이 활발하게 연구[3,6-8]되고 있으며, 분석 대상 데이터들을 스케일링(Scaling)이라고 하는 일종의 해석기준에 따라서 이진데이터로 변환하여 개념 추출 및 개념계층구조를 생성하고 있다. 주어진 해석기준이 얼마나 정교/적합한가에 따라서 다양한 형식개념분석기법 결과(개념계층구조, 연관규칙)를 얻을 수 있다.

한편 형식개념분석기법에서는 공통속성을 갖는 객체들을 추출한다는 전제조건을 토대로 개념을 추출하고 개념계층구조를 구축하고 있다. 따라서, 다양한 상황이나 조건에 적합한 “공통성(commonality)”이외의 다른 조건이나 제약사항들을 기반으로 하는 새로운 개념들을 추출하기에는 한계가 있다. 이와 같은 문제를 해결하기 위한 한 가지 방법으로써, 본 논문에서는 입자화 정도(granularity level)를 기반으로 하는 형식개념분석기법을 제안한다. 본 논문에서 제안하는 기법에서는 형식개념분석기법에 입자화 정도를 도입하여 다양한 조건과 추상화 수준을 토대로 하여, 개념들을 추출하고 개념계층구조를 구축할 수 있다.

본 논문은 다음과 같이 구성된다. 제2장에서는 형식개념분석기법 관련연구에 대해서 설명하고 Granular Computing 분야에서 입자화 정도에 관한 연구에 대하여 설명한다. 제3장에서는 분석 대상 객체들의 속성에 대한 입자화 정도를 나타낸 입자화 정도 트리(gl-tree :granularity level tree)에 대한 정의와 이를 토대하는 입자개념계층구조(Granular Concept Hierarchy)의 구축에 대하여 설명한다. 제4장에서는 결론과 향후 연구과제에 대

해서 설명한다.

2. 관련연구

여기서는 본 연구의 관련연구로서, 형식개념분석기법의 기본 개념[3]과 Granular Computing 분야에서 입자화 정도에 관한 연구[1, 9]에 대하여 설명한다.

형식개념분석기법에서는 주어진 도메인으로부터 객체(Object)와 속성(Attribute)들을 추출하고, 이들 사이의 포함관계를 이진데이터 테이블 형태로 표현한다. 이진데이터 테이블 K = (G, M, I)는 객체들(Objects)의 집합 G와 속성들(Attributes)의 집합 M, 그리고 G와 M사이의 이항관계 I⊆G×M로 구성된다. 즉, 어떤 객체 g가 속성 m을 가지고 있을 경우, gIm 또는 (g, m) ∈ I 로 나타낸다.

이진데이터 테이블 에 대하여, O⊆G, A⊆M 일 때, O′=A ∧ A′=O를 만족하는 (O, A)를 개념(Concept)이라고 한다. 단, O′ := {a∈M|∀o∈O:(o, a)∈I}, A′ := {o∈G|∀a∈A:(o, a)∈I}. 즉, 개념 (O, A)는 O의 모든 객체들이 공통적으로 갖는 속성들의 집합이 A와 같고, A의 모든 속성들을 공통적으로 갖는 객체들의 집합이 O와 같은 객체집합과 속성집합으로 구성된다. 또한, 임의의 개념 (O1, A1), (O2, A2)에 대하여, O1⊆O2 (⇔A1⊇A2)라면, (O1, A1)은 (O2, A2)의 상위개념(또는, (O2, A2)는 (O1, A1)의 하위개념)이며, (O1, A1)≤(O2, A2)와 같이 표현한다. 이진데이터 테이블 K=(G, M, I)로부터 만들어진 모든 개념들의 집합 C와 그들 사이의 상·하위개념관계≤로 이루어진 계층구조 L:=(C, ≤)을 개념계층구조(Concept Lattice)라고 부른다.

Granular Computing을 기반으로 하는 데이터 분석 분야에서 파티션은 대상이 되는 객체집합 U를 입자화하는데 사용되는 가장 일반적인 방법이다. 객체들의 집합 U에 대한 파티션 π = {X_i | 1≤ i ≤n}는 공통 원소를 갖지 않는 U의 부분집합 n개로 구성된 집합으로, 공집합을 원소로 갖지 않으며, 파티션 π의 모든 원소를 합집합하면 전체집합 U가 나오게 되는 특징을 갖고 있다. 대상이 되는 객체들의 집합 U에 대한 다양한 파티션들이 생성될 수 있으며, 그 중 파티션들 사이에 상•하위파티션관계(≤)가 형성되는 파티션들이 존재할 수 있다. 상•하위파티션관계는 하위 파티션의 입자들이 상위 파티션의 입자들에 모두 포함되는 것을 의미한다. 예를 들어, 객체들의 집합 U= {o1, o2, o3, o4}에 대한파티션들을 모두 구하면 다음과 같다. 15개의 파티션 중에서 파티션 π_{13} ={{o3}, {o1 ,o2}, {o4}}과 π_5 = {{o1, o2}, {o3, o4}}는 상•하위파티션관계이고 $\pi_{13}\leq\pi_5$와 같이 나타낸다. 즉, 하위 파티션 π_{13}의 모든 입자들이 상위 파티션 π_5의 입자들의 부분집합(⊆)임을 나타낸다.

Partition	
π_0	{{o1, o2, o3, o4}}
π_1	{{o1}, {o2, o3, o4}}
π_2	{{o2}, {o1, o3, o4}}
π_3	{{o3}, {o1, o2, o4}}
π_4	{{o4}, {o1, o2, o3}}
π_5	{{o1, o2}, {o3, o4}}
π_6	{{o1, o3}, {o2, o4}}
π_7	{{o1, o4}, {o2, o3}}
π_8	{{o1}, {o2, o3}, {o4}}
π_9	{{o1}, {o2, o4}, {o3}}
π_{10}	{{o1}, {o3, o4}, {o2}}
π_{11}	{{o2}, {o1, o3}, {o4}}
π_{12}	{{o2}, {o1, o4}, {o3}}
π_{13}	{{o3}, {o1, o2}, {o4}}
π_{14}	{{o1}, {o2}, {o3}, {o4}}

이와 같이, 대상이 되는 객체들의 집합 U에 대해 입자들을 추출하고 상•하위파티션관계를 갖는 파티션들을 만들기 위해서는 적당한 기준이 있어야 한다. 관련연구[9]에서는 속성에 대한 입자화 정도 트리(gl-tree)를 구축하여 추출되는 입자들의 정도를 조절하는 방법을 제안하고 있다.

3. Granular Concept Hierarchy 구성

본 장에서는 분석 대상 객체들의 입자화 정도를 나타내는 입자화 정도 트리(gl-tree) 구축을 위한 제반 정의들과 gl-tree를 기반으로 하는 입자개념계층구조(Granular Concept Hierarchy)의 구축에 대하여 설명한다.

형식개념분석기법과 같은 데이터분석기법에서 분석대상이 되는 데이터는 여러 가지 다양한 값을 가지는 속성들과 객체들, 그리고 객체와 속성 사이의 관계를 나타내는 테이블형태로 정리되어 표현할 수 있다[1-3].

[정의1] 데이터 테이블 S = (U, A, V, I)는 다음과 같은 요소들로 구성된다:

- 객체들의 집합 U = {u1, u2, …, u_m},
- 속성들의 집합 A = {a1, a2, …, a_n},
- 속성이 갖는 값들의 집합 V = {V_{a1}, V_{a2}, …, V_{an}},
- 함수 I(u, a_i) : U×A → V_{ai}는 다음의 조건을 만족하는 U×A로부터 V_{ai}로의 사상:

$$I(u, a_i)\in V_{ai},\ \forall u\in U,\ a_i\in A\blacksquare$$

즉, U와 A의 원소는 각각 해당 테이블의 객체들과 각 객체들이 가질 수 있는 속성들, 그리고 그 속성의 값들을 나타낸다. 또한, I(u, a_i):U×A → V_{ai}는 어떤 객체 u가 속성 a_i를 가지고 있고 그 속성의 값이 v∈Va_i임을 나타낸다.

표1은 자동차에 대한 데이터 테이블로써, 5대의 자동차와 그 자동차들의 "제조사", "배기량", "색상"에 대한 정보를 나타낸다.

<표 1> 자동차에 대한 데이터 테이블

객체 \ 속성	제조사	배기량	색상
O1	기아	1591	red
O2	아우디	2967	black
O3	렉서스	4608	white
O4	현대	995	pink
O5	BMW	2979	grey

이와 같은 다양한 값을 갖는 데이터 테이블로부터 개념들을 추출하고 개념계층구조를 구축하기 위해서는 특정한 규칙에 따라 주어진 데이터 테이블을 이진데이터 테이블로 변환할 필요가 있다. 이러한 변환과정을 스케일링(Scaling)이라고 한다. 스케일링하기 위해서, 주어진 데이터 테이블의 각 속성들은 스케일 테이블(해석기준)을 토대로 이진데이터 테이블로 변환되며, 다양한 값을 갖는 데이터 테이블의 각 속성들은 스케일 테이블에 의해서 해석된다.

[정의2] 다양한 값을 갖는 데이터 테이블 S의 속성 m∈A에 대한 스케일 테이블 Sm = (Vm, Am, Im)은 데이터 테이블 S에서 속성 m이 갖는 값들의 집합 Vm과 속성m

을 세분화하여 표현한 속성들의 집합 Am, 그리고 Vm과 Am사이의 이항관계 Im⊆Vm×Am로 구성된다[3].■

표1에서 속성 "배기량"에 대한 스케일 테이블 $S_{배기량}$은 표2와 같다.

<표 2> 속성"배기량"에 대한 스케일 테이블

$S_{배기량}$	소형	중형	대형
1591	X		
2967		X	
4608			X
995	X		
2979		X	

스케일 테이블은 하나의 속성을 대상으로 다양한 관점으로 세분화하여 여러 개의 스케일 테이블들을 만들어 낼 수 있다. 이와 같이 다양한 값을 갖는 데이터 테이블과 데이터를 해석하는 기준이 되는 스케일 테이블을 토대로 입자화 정도 트리를 다음과 같이 구성할 수 있다.

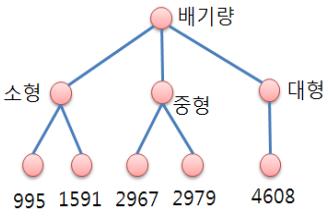
[정의3] 데이터 테이블 S = (U, A, V, I)와 임의의 속성 m∈A에 대한 스케일 테이블 Sm = (Vm, Am, Im)이 주어졌을 때, 속성 m에 대한 입자화 정도 트리 GLT(m) 은 다음과 같이 정점m을 루트(root)로 하는 트리로 정의한다:

GLT(m) = (V, E)

단, V = {m}∪Am∪Vm, E는 다음과 같은 방식으로 구성되는 변들의 집합이다.

(1) E = E∪{(m, x) | ∀x∈Am}

(2) E = E∪{(m, y) | ∃x∈Am, y∈Vm : (x, y)∈Im}■



(그림1) 표1의 속성 "배기량"에 대한 입자화 정도 트리

그림1은 자동차에 대한 데이터 테이블(표1)과 속성 "배기량"에 대한 스케일 테이블(표2)이 주어졌을 때, 속성 "배기량"에 대한 입자화 정도 트리이다.

속성 m에 대한 입자화 정도 트리 GLT(m) = (V, E)가 주어졌을 때, gLevel은 GLT(m)에서의 깊이(depth)를 나타낸다.

[정의4] 속성 m에 대한 입자화 정도 트리 GLT(m) = (V, E)가 주어졌을 때, gLevel l에 존재하는 정점들의 집합 GLV(m, l)은 다음과 같이 정의 한다:

GLV(m, l) = {x∈V | depth(x) = l}.

단, depth(x)는 정점 x의 깊이이다.■

그림1과 같이 속성 "배기량"에 대한 입자화 정도 트리에서 gLevel 0, 1, 2 각각에 대한 GLV는 다음과 같다.

(1) GLV(배기량, 0) = {배기량}

(2) GLV(배기량, 1) = {소형, 중형, 대형}

(3) GLV(배기량, 2) = {1591, 2967, 4608, 995, 2979}.

[정의5] 주어진 GLT(m)에 대하여, 임의의 gLevel l의 정점집합 GLV(m, l)을 기준으로 하는 파티션 π(m, l)은 다음과 같이 정의한다.

$$\pi(m, l) = \{X_i \mid 1 \le i \le n\} = \bigcup_{x \in GLV(m,l)} \{x' \in U\}$$

단, 다음 조건을 만족한다.

(1) $\forall i,\ X_i \neq \varnothing$

(2) $\forall i \neq j,\ X_i \cap X_j \neq \varnothing$

(3) $\bigcup \{X_i \mid 1 \le i \le n\} = U$ ■

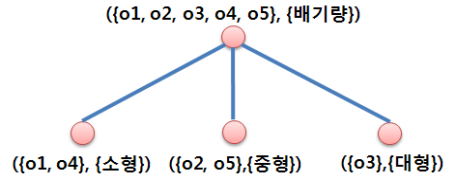
표1과 같은 데이터 테이블의 속성 "배기량"에 대한 입자화 정도 트리(그림1)를 토대로 다음과 같이 3개의 파티션을 추출할 수 있으며, 그들 사이에는 상•하위파티션관계가 존재한다.

(1) 파티션 π(배기량, 0) = {{o1, o2, o3, o4, o5}}

(2) 파티션 π(배기량, 1) = {{o1, o4}, {o2, o5}, {o3}}

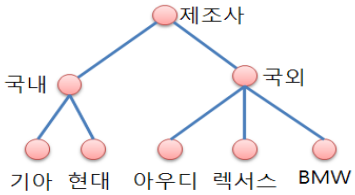
(3) 파티션 π(배기량, 2) ={{o1}, {o2}, {o3}, {o4}, {o5}}

GLT(배기량)에 대한 GLV와 파티션을 기반으로 입자개념(granular concept)들을 생성할 수 있다. 입자개념은 파티션을 구성하는 입자와 GLV의 원소를 하나의 쌍(pair)으로 묶어놓은 것을 의미하며, 생성된 입자개념들 사이에는 입자들 사이의 부분집합 관계에 의해 상•하위입자개념 관계가 형성된다. 입자개념들과 그들 사이의 상•하위입자개념 관계를 토대로 다음과 같이 속성 "배기량"에 대한 입자개념계층구조(Granular Concept Hierarchy)를 구성할 수 있다.

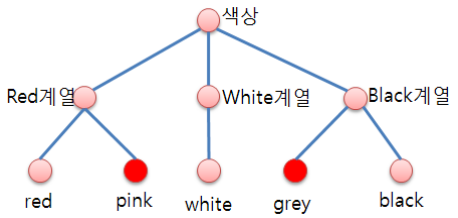


(그림 2) 속성 "배기량"에 대한 입자개념개층구조

앞서 설명한 정의들을 토대로 표1의 속성 "제조사"와 "색상"에 대한 입자화 정도 트리를 다음과 같이 각각 생성할 수 있다(그림3, 4 참조).



(그림 3) 표1의 속성 "제조사"에 대한 입자화 정도 트리



(그림 4) 표1의 속성 "색상"에 대한 입자화 정도 트리

구축된 입자화 정도 트리를 토대로, gLevel에 따른 GLV와 파티션들을 다음과 같이 추출할 수 있다.

(1) GLT(제조사)에 대한 GVL과 파티션 추출 결과

-GLV(제조사, 0) = {제조사}

-GLV(제조사, 1) = {국내, 국외}

-GLV(제조사, 2)={기아, 현대, 아우디, 렉서스, BMW}

- 파티션 π(제조사, 0) = {{o1, o2, o3, o4, o5}}

- 파티션 π(제조사, 1) = {{o1, o4}, {o2, o3, o5}}

- 파티션 π(제조사, 2) ={{o1}, {o2}, {o3}, {o4}, {o5}}

(2) GLT(색상)에 대한 GVL과 파티션 추출 결과

-GLV(색상, 0) = {색상}

-GLV(색상, 1) = {Red계열, White계열, Black계열}

-GLV(색상, 2)={red, pink, white, black, grey}

- 파티션 π(제조사, 0) = {{o1, o2, o3, o4, o5}}

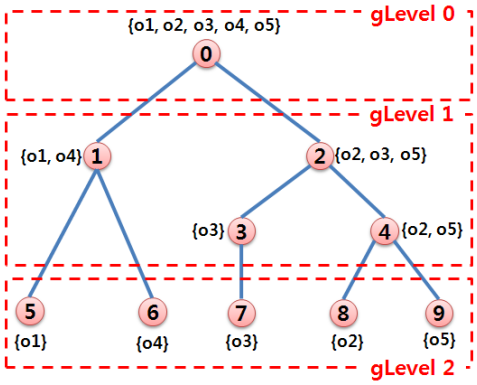
- 파티션 π(제조사, 1) = {{o1, o4}, {o3}, {o2, o5}}

- 파티션 π(제조사, 2) ={{o1}, {o2}, {o3}, {o4}, {o5}}

추출된 GLV와 파티션들을 토대로 생성된 입자개념들(표3)과 그들 사이의 상•하위입자개념 관계를 기반으로 gLevel에 따라서 각 속성에 대한 다양한 입자화정도를 갖는 입자개념계층구조(그림5)를 구축할 수 있다.

<표 3> 표1로부터 추출된 모든 입자개념들

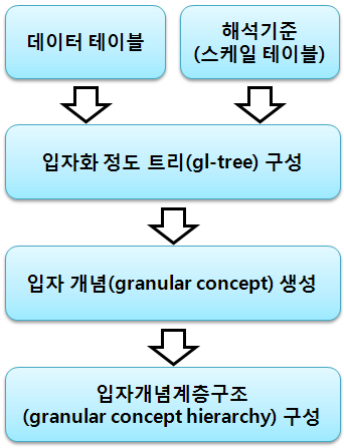
	입자(granular)	의미(meaning)
0	{o1, o2, o3, o4, o5}	{제조사, 배기량, 색상}
1	{o1, o4}	{제조사_국내, 배기량_소형, 색상_Red계열}
2	{o2, o3, o5}	{제조사_국외}
3	{o3}	{배기량_대형, 색상_White계열}
4	{o2, o5}	{배기량_중형, 색상_Black계열}
5	{o1}	{제조사_기아, 배기량_1591, 색상_red}
6	{o4}	{제조사_현대, 배기량_995, 색상_pink}
7	{o3}	{제조사_렉서스, 배기량_4608, 색상_white}
8	{o2}	{제조사_아우디, 배기량_2967, 색상_black}
9	{o5}	{제조사_BMW, 배기량_2927, 색상_grey}



(그림 5) 표1에 대한 입자개념계층구조

4. 결론

본 논문에서는 데이터 테이블의 해석을 위한 기준으로 각 속성에 대한 스케일 테이블을 토대로 입자화 정도 트리(gl-tree)를 구축하고 이를 기반으로 입자개념들을 생성하여, 입자개념들 사이에 관계를 토대로 입자개념계층구조를 구축하기 위한 방법을 제안하였다(그림6). 본 논문에서 제안한 기법을 토대로, 각 속성에 대하여 다양한 상황과 조건, 그리고 추상화 수준을 적용한 입자개념 및 입자개념계층구조를 생성하고 가시화할 수 있다. 또한, 구축된 입자개념계층구조를 토대로 속성들 사이의 연관규칙을 추출하여 사용자로부터의 질의에 대한 추론규칙을 생성할 수 있는 토대를 제공할 수 있다.



(그림 6) 입자개념계층구조의 구축

본 논문의 연구에서는, 데이터 분석 대상이 되는 객체 집합으로부터 입자를 추출하기 위한 기준으로서 파티션(partition)을 이용하였으나, 보다 다양한 형태의 입자를 추출하기 위하여 커버링(Covering)을 적용한 입자화 및 입자개념계층구조의 구축기법에 대한 연구, 그리고 본 논문에서 제안한 기법들을 자동화하기 위한 도구의 개발 등이 필요하다.

참고문헌

[1] J.T. Yao, Y. Y. Yao and Y. Zhao, "Foundations of Classification",

[2] B. Zhou, Y. Yao, "A Logic Approach to Granular Computing", International Journal of Cognitive Informatics and Natural Intelligence, Vol 2, Issue 2, pp. 1-28, 2008.

[3] B. Ganter, R. Wille, Formal Concept Analysis: Mathematical Foundations, Springer, 1999.

[4] Z. Pawlak, Rough Sets : Theoretical Aspects of Reasoning about Data, Springer, 1991.

[5] R. Lowen, Fuzzy Set Theory: Basic Concepts, Techniques and Bibliography, Springer, 1996

[6] Susanne P., "Logical Scaling in Formal Concept Analysis", ICCS, pp.332-341, 1997.

[7] 강유경, 황석형, 최희철, 김동순, 김홍기, 김명기, "Many-valued Context의 Scaling을 위한 형식개념분석 도구의 개발", 한국정보처리학회 춘계학술발표대회 논문집, 제12권, 제2호, pp. 251-254, 2005.

[8] Belohlavek R., Konecny J. "Scaling, Granulation, and Fuzzy Attributes in Formal Concept Analysis", The IEEE International Conference on Fuzzy Systems, pp. 918-923, 2007.

[9] R. Belohlav다, V. Sklenar, "formal concept analysis over attributes with levels of granularity", CIMCA-IAWTIC'05, pp.619-624, 2005.