러프 및 퍼지 데이터의 형식개념분석을 지원하기 위한 도구의 개발

강유경*, 황석형*, 김응희**
*선문대학교 컴퓨터공학부, **서울대학교 치과대학
e-mail: {aquamint99, shwang}@sunmoon.ac.kr, eungheekim@snu.ac.kr

Development of tools to support Formal Concept Analysis for Rough and Fuzzy Data

Yu-Kyung Kang*, Suk-Hyung Hwang*, Eung-Hee Kim**
*Dept of Computer Science & Engineering, SunMoon University
**School of Dentistry, Seoul National University

요 약

실세계의 복잡하고 다양한 데이터에 내포된 유용한 정보들을 추출하여 활용하기 위해 다양한 데이터 마이닝 기법들이 제안되고 있다. 최근 각광받기 시작한 개념분석기법(Formal Concept Analysis)은, 주어진 데이터로부터 개념들을 추출하고 그들 사이의 관계를 파악하여 개념계층구조를 구축하기 위한 정형화된 데이터분석 기법이다. 본 논문에서는 개념분석기법을 기반으로 다종다양한 데이터를 분석할 수 있는 기법들(FFCA, RFCA)에 대해서 소개하고, 본 연구에서 개발하고 있는 지원도구와 그 도구를 이용한 실험 결과를 보고한다.

1. 서론

오늘날, World Wide Web의 발달로 다종다양한 컴퓨팅 환경에서 매일매일 방대한 양의 데이터들과 지식, 정보들이 기하급수적으로 발생되고 유통되면서, 우리는 정보의 홍수시대를 살아가고 있다. 따라서, 실세계의 복잡하고 다양한 데이터에 내포된 유용한 정보들을 추출하여 활용하기 위해 다양한 데이터 마이닝(DM : Data Mining) 기법들이 필요하다[1, 2]. 대량의 데이터로부터 의미 있는 새로운 정보를 추출하기 위한 데이터 마이닝 기법으로는 데이터 사이의 관련성을 표현하는 연관성 측정(Associations)과 과거에 수집된 데이터들을 분석하여 이들 사이에 존재하는 패턴을 트리형태로 만들어서 새로운 데이터를 분류하고 예측하는 의사결정 트리(Decision trees)가 있다. 또한, 신경망 모형(Neural networks)은 인간 두뇌의 신경세포를 모방한 개념으로 인간의 경험으로부터 학습해 가는 두뇌의 신경망 활동을 흉내 내어 주어진 데이터에 대해 반복적인 학습과정을 거쳐 패턴을 찾아내고 이를 일반화함으로써, 향후를 예측하고자 하는 데이터 마이닝 기법이다. 이 이외에도 전통적인 통계기법(Statistical analysis) 및 순차적패턴발견(Sequential Pattern discovery), 분류(Classification), 요약(Summarization), 군집화(Clustering) 등 다양한 분석기법들이 있다[1-5].

최근 주목받고 있는 개념분석(Concept Analysis)기법은 주어진 도메인 내의 데이터들로부터 개념들(Concepts)을 추출하고, 개념들 사이의 상-하위관계(Super-Sub Relation)를 파악하여 개념계층구조(Conceptual hierarchy)를 구축하기위한 정형화된 데이터 마이닝의 일종이다. 개념분석기법을 사용함으로써 실세계의 데이터에 함축된 개념들에 대한 계층구조를 효과적으로 구축할 수 있기 때문에, 현재 의학, 정보과학, 소프트웨어 공학, 등 다양한 분야에서 적용하여 활용되고 있다[5-10].

본 논문에서는, 실세계에 존재하는 다양한 데이터로부터 유용한 정보를 추출하기 위한 데이터 마이닝 기법으로써, 개념분석 기법을 기반으로 하는 FCA(Formal Concept Analysis)와 FFCA(Fuzzy Formal Concept Analysis), RFCA(Rough Formal Concept Analysis)에 대해서 소개하고, 본 연구에서 개발하고 있는 지원도구와 그 도구를 이용한 실험 결과를 보고한다. 본 논문은 다음과 같이 구성된다. 2장에서는 FCA 및 FFCA와 RFCA의 기본적인 수학적 정의에 대해 설명한다. 그리고 3장에서는 각 기법을 지원하기 위한 도구로써, 본 연구에서 개발하고 있는 Concept Analyzer에 대해서 소개한다. 4장에서는 Concept Analyzer를 이용하여 실시한 실험결과에 대해 보고하고, 결론 및 향후 연구과제에 대해서 설명한다.

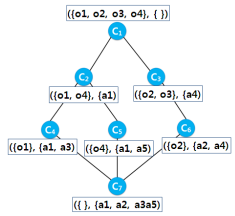
2. 개념분석 기법

2.1. FCA(Formal Concept Analysis)

FCA의 기본이 되는 Formal context 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 나타내며, g는 m을 갖는다는 것을 의미한다. Formal context는 cross table형태로 나타낼 수 있으며, 해당 표의 행과 열의 헤드부분은 각각 context를 구성하는 객체들과 속성들로 구성된다. 해당 셀에 관련된 객체와

<표 1> Formal Context

	a1	a2	a3	a4	a5
o1	X		X		
o2		X		X	
o3				X	
o4	X				X



(그림 1)Concept Lattice

속성이 이항관계 I를 만족할 경우에는 X표시하고, 이외의 경우에는 빈 공란으로 남겨둔다(표1 참조).

이때, Formal context K=(G, M, I)에 대하여, O⊆G, A⊆M 일 때, O′=A ∧ A′=O를 만족하는 (O, A)를 개념(formal concept)이라고 한다. 단, O′ := {a∈M|∀o∈O:(o, a)∈I}, A′ := {o∈G|∀a∈A:(o, a)∈I}. 즉, formal concept (O, A)는 O의 모든 객체들이 공통적으로 갖는 속성들의 집합이 A와 같고, A의 모든 속성들을 공통적으로 갖는 객체들의 집합이 O와 같음을 의미한다. 또한, 임의의 개념 (O_1, A_1), (O_2, A_2)에 대하여, $O_1 \subseteq O_2$ $(\Leftrightarrow A_1 \supseteq A_2)$라면, (O_1, A_1)은 (O_2, A_2)의 상위개념(또는, (O_2, A_2)는 (O_1, A_1)의 하위개념)이며, $(O_1, A_1) \leq (O_2, A_2)$와 같이 표현한다.

Formal context K=(G, M, I)로부터 만들어진 모든 개념들과 그들 사이의 상위-하위개념관계≤로 이루어진 계층구조를 개념격자(Concept Lattice 또는 Galois Lattice)라고 부른다(그림 1 참조).

개념격자를 나타낸 Hasse Diagram에서는, 각 개념들과 이들 사이의 상하위관계가 링크에 의해 표시되며, 특히, 개념들 간의 링크에 의해 만들어지는 경로에 의해 상위개념으로부터 하위개념으로 속성들이 상속되며, 하위개념으로부터 상위개념으로 해당 객체들이 전파된다. FCA에서는, 주어진 문제영역의 객체들과 이들이 갖는 속성들을 context형태로 파악하여, 개념을 추출하고 개념격자형태로 나타냄으로써, 도메인 내의 개념들을 분류하고 체계화 할 수 있는 계층적 개념구조를 구축할 수 있다.

2.2. FFCA(Fuzzy Formal Concept Analysis)

실세계에 존재하는 애매모호한 데이터를 다루기 위해 퍼지집합이론(FST : Fuzzy Set Theory)[11, 12]과 러프집합이론(RST : Rough Set Theory)[13-16]이 제안되었다. 퍼지집합은 집합의 각 원소가 그 집합에 귀속되어지는 정도를 [0, 1]내의 수로 나타낸 귀속도(membership degree)를 갖는 원소들의 집합이다. 이와 같은 퍼지집합에 FCA를 접목함으로써, 주어진 데이터로부터 숨겨져 있는 지식을 귀속도를 포함한 개념단위로 추출하여 구조화 할 수 있다.

[정의 1] *Fuzzy context* K := (G, M, I=φ (G×M))는 객체들의 집합 G와 속성들의 집합 M, 그리고 G와 M사이의 관계를 나타내는 I로 구성된다. 단, (g, m)∈ I는 0과 1사이의 귀속도μ (g, m)를 갖는다.■

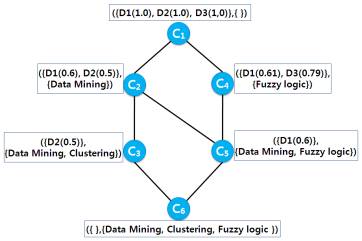
표 2은 Fuzzy context의 예로써, 3개의 문서에 "Data Mining", "Clustering", "Fuzzy logic"이라는 용어의 출현빈도수를 나타낸 것이다. 예를 들어, 문서 D1에는 "Data Mining"이라는 용어가 0.6정도 출현한다. 귀속도가 낮은 데이터를 제거하기 위해 주어진 도메인에 맞게 임계값(Threshold) T를 사용자가 지정할 수 있다. 표 3은 표2에 T= 0.5를 적용한 결과이다.

<표 2> Fuzzy context

	Data Mining	Clustering	Fuzzy logic
D1	0.6	0.12	0.61
D2	0.5	0.7	0.47
D3	0.2	0.3	0.79

<표 3> 표 2에 임계값 T = 0.5를 적용한 결과

	Data Mining	Clustering	Fuzzy logic
D1	0.6	-	0.61
D2	0.5	0.7	-
D3	-	-	0.79



(그림 2) Fuzzy Concept Lattice

[정의 2] 임의의 Fuzzy context K와 임계값 T에 대하여, A⊆G, B⊆M일 때, $A^* = B \wedge B^* = A$를 만족하는 (φ (A), B)를 *Fuzzy concept*이라고 한다. 단, A^*={m∈M | ∀g∈A : μ (g, m)≥T}, B^*= {g∈G | ∀m∈B : μ(g, m) ≥T}. 또한, 각 객체 g∈φ (A)에 대해서 $\mu_g = \min_{m \in B} \mu(g, m)$.■

즉, μ_g는 객체 g와 속성 m 사이의 귀속도 중 최소값을 나타낸다. 만약, B = {} 라면 모든 g∈A에 대한 μ_g = 1이다. 표 2로부터 추출된 모든 Fuzzy concept들은 표4와 같다(T=0.5). 예를 들어, 표3에 대해서, A = {D2}이고, B = {Data Mining, Clustering}일 때, $D2^*$ = {Data Mining, Clustering}이고, ({Data Mining, Clustering})* = {D2}이므로 ({D2}, {Data Mining, Clustering})는 Fuzzy concept이다. 단, μ (D2, Data Mining) = 0.5이고, μ (D2, Clustering) = 0.7이므로, μ_{D2} = 0.5이다. 따라서, ({D2(0.5)}, {Data Mining, Clustering})와 같이 표현된다.

<표 4> 표2에 대한 Fuzzy concepts

ID	Extents	Intents
C_1	{D1(1.0), D2(1.0), D3(1.0)}	{ }
C_2	{D1(0.6), D2(0.5)}	{Data Mining}
C_3	{D2(0.5)}	{Data Mining, Clustering}
C_4	{D1(0.61), D3(0.79)}	{Fuzzy logic}
C_5	{D1(0.6)}	{Data Mining, Fuzzy logic}
C_6	{ }	{Data Mining, Clustering, Fuzzy logic}

Fuzzy context K로부터 추출된 모든 Fuzzy concept들 사이에는 다음과 같은 상위-하위개념관계(Super-sub concept relation)가 존재한다.

[정의 3] 임의의 Fuzzy concept C_1 = (φ (A_1), B_1), C_2 = (φ (A_2), B_2)에 대하여, *상위-하위개념관계* (φ (A_1), B_1) ≤ (φ (A_2), B_2)는 다음과 같이 정의 된다.

(φ (A_1), B_1)≤(φ (A_2), B_2) ⇔ φ (A_1)⊆φ (A_2)(⇔$B_1 \supseteq B_2$).■

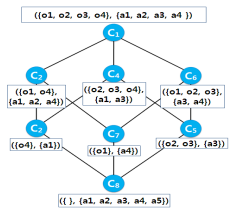
Fuzzy concept C_2 = ({D1(0.6), D2(0.5)}, {Data Mining}) 와 C_3 = ({D2(0.5)}, {Data Mining, Clustering})에 대해서, {D2(0.5)}⊆{D1(0.6), D2(0.5)}(⇔{Data Mining, Clustering}⊆{Data Mining})이므로, C_2는 C_3의 상위개념(Super Concept)이며, $C_2 \leq C_3$와 같이 표현한다.

임의의 Fuzzy context K에 대한 Fuzzy Concept Lattice L은 K에 대한 임계값 T와 K로부터 추출된 모든 Fuzzy concept들과 그들 사이의 상위-하위개념관계들에 의해 그림 2와 같이 표현된다.

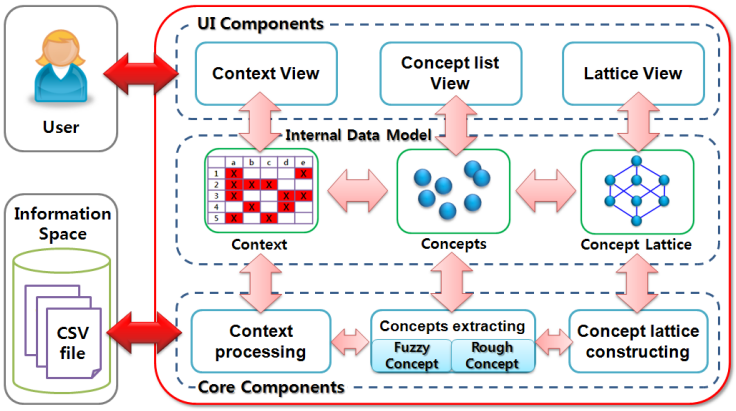
2.3. RFCA(Rough Formal Concept Analysis)

<표 5> Formal context

	a1	a2	a3	a4
o1		X		X
o2			X	
o3			X	
o4	X	X		



(그림 3) Rough Concept Lattice



(그림 4) Concept Analyzer Architecture

RFCA는 애매모호한 데이터를 다루기 위한 러프집합 이론에 FCA를 접목하여 동치관계를 기반으로 구별하기 애매모호한 원소들을 포함시켜서 하나의 개념으로 추출하여 주어진 데이터를 분류하기 위한 개념분석 기법이다.

RFCA의 입력 데이터는 표 5와 같은 Formal context이다.

러프집합이론은 애매하고 모호한 데이터를 다루기 위해서 다음과 같은 중요한 함수들을 제공한다.

[정의 4] 임의의 Formal context K := (G, M, I)에 대해서, X⊆G, Y⊆M 일 때, *하한근사(LA)*와 *상한근사(UA)*는 다음과 같이 정의한다.

$$LA(X) = \{a \in M \mid OS(a) \subseteq X\},$$
$$UA(X) = \{a \in M \mid OS(a) \cap X \neq \varnothing\}.$$
$$LA(Y) = \{o \in G \mid AS(o) \subseteq Y\},$$
$$UA(Y) = \{o \in G \mid AS(o) \cap Y \neq \varnothing\}.$$

단, $OS(a) := \{o \in G \mid (o, a) \in I\}$, $AS(o) := \{a \in M \mid (o, a) \in I\}$.■

LA(X)는 X가 배타적으로 갖는 속성들의 집합이고, UA(X)는 X가 총체적으로 갖는 속성들의 집합이다. 예를 들어, 표 5에 대해서, X = {o1, o2, o3} 일 때, LA(X) = {a3, a4}이고, UA(X) = {a2, a3, a4}이다.

LA와 UA함수를 사용하여 주어진 Formal context로부터 Rough concept을 추출 할 수 있다.

[정의 5] 임의의 Formal context K := (G, M, I)에 대해서, X⊆G, Y⊆M 일 때, X = UA(Y) ∧ Y = LA(X)를 만족하는 (X, Y)를 *Rough concept*이라고 한다.■

예를 들어, 표 5에 대해서, X = {o1, o2, o3} 일 때, LA(X) = {a3, a4}이고, UA(LA(X)) = UA({a3, a4}) = {o1, o2, o3}이다. 즉, X = UA(Y) ∧ Y = LA(X)이므로, ({o1, o2, o3}, {a3, a4})는 Rough concept이다. 이와 같은 방법으로 표 5로부터 8개의 Rough concept들을 추출할 수 있다(표6 참조).

[정의 5]에 의해서 추출된 Rough concept들 사이에는 "sub-concept of"관계(≤)가 존재한다.

<표 6> 표 5에 대한 Rough concepts

ID	Extents	Intents
C_1	{o1, o2, o3, o4}	{a1, a2, a3, a4}
C_2	{o4}	{a1}
C_3	{o1, o4}	{a1, a2, a4}
C_4	{o2, o3, o4}	{a1, a3}
C_5	{o2, o3}	{a3}
C_6	{o1, o2, o3}	{a3, a4}
C_7	{o1}	{a4}
C_8	{ }	{ }

[정의 6] 임의의 Rough concept $C_1 = (X_1, Y_1)$, $C_2 = (X_2, Y_2)$에 대하여, "sub-concept of"관계 $(X_1, Y_1) \lesssim (X_2, Y_2)$는 다음과 같이 정의 된다.

$$(X_1, Y_1) \lesssim (X_2, Y_2) \Leftrightarrow X_1 \subseteq X_2 (\Leftrightarrow Y_1 \subseteq Y_2).$$■

Rough concept C_5 = ({o2, o3}, {a3})와 C_6 = ({o1, o2, o3}, {a3, a4})에 대해서, {o2, o3}⊆{o1, o2, o3}(⇔{a3}⊆{a3, a4})이므로, C_5는 C_6의 하위개념(Sub concept)이며, $C_5 \lesssim C_6$와 같이 표현한다.

위의 정의들을 토대로 Rough Concept Lattice는 그림3과 같다.

3. Concept Analyzer

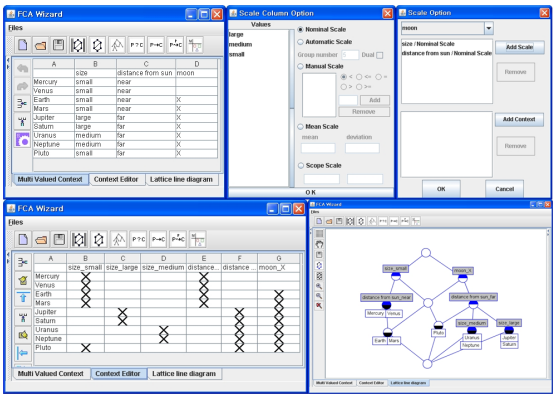
앞 절의 제반 정의들을 토대로, 각 분석기법을 지원하기 위한 Concept Analyzer를 개발하였으며, 전체적인 아키텍처는 그림 4와 같다.

3.1. Core Components

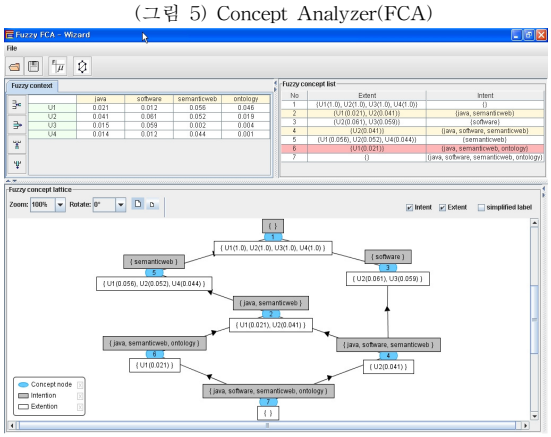
- Context processing component : 이 컴포넌트는 Information Space로부터 데이터를 읽어드려서 formal context 형태로 표현한다.
- Concepts extracting component : 이 컴포넌트는 주어진 formal context로부터 개념 및 그들 사이의 관계들을 추출하는 컴포넌트로써, formal concept extracting, fuzzy concept extracting, 그리고 rough concept extracting 모듈들을 포함하고 있다.
- Concept lattice constructing component : Concepts extracting component로부터 추출된 개념들과 그 개념들 사이의 상-하위 관계를 파악하여 concept lattice를 구축하는 컴포넌트이다.

3.2. UI Components

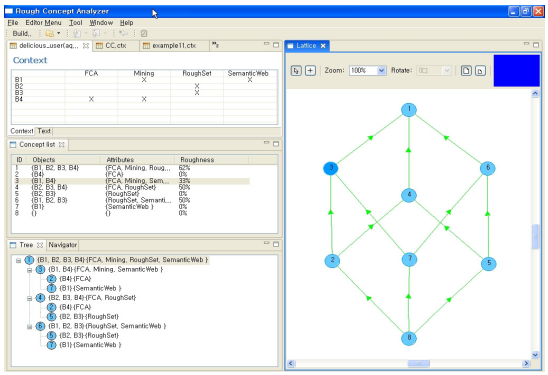
- Context View component : Context processing component에 의해 Information Space로부터 읽어드린 데이터를 사용자에게 formal context 형태로 나타낸다. 또한, context 생성 및 편집 기능도 제공한다.
- Concept list View component : 이 컴포넌트는 Concepts extracting component에서 추출된 모든 개념들에 대한 정보를 사용자에게 표 4와 같은 형태로 보여준다.
- Lattice View component : 이 컴포넌트는 Concept lattice constructing component에 의해 구축된 Lattice를 그래픽하게 가시화하는 컴포넌트이다.



(그림 5) Concept Analyzer(FCA)



(그림 6) Concept Analyzer(FFCA)



(그림 7) Concept Analyzer(RFCA)

4. 결론 및 향후 연구 과제

본 논문에서는 실세계에 존재하는 다양한 데이터로부터 유용한 정보를 수월하게 추출하기 위한 FCA와 FFCA, RFCA를 소개하고, 이를 기반으로 하는 Concept Analyzer를 개발하였다. 본 연구결과의 유용성을 살펴보기 위하여, Concept Analyzer를 이용하여 실제 Folksonomy 시스템에서 발생하는 데이터(어떤 사용자가 어떤 태그를 사용하는지)에 FCA를 적용한 결과, 공통 태그를 사용하는 사용자들을 Concept 단위로 추출하여 구조화 하였으며, 구조화된 Concept lattice를 통해서 이 사용자들이 어떤 태그들을 공통으로 사용하고 있는지, 어떤 태그를 사용하는 사용자들(Concept)과 관계가 있는지 파악할 수 있었다. 또한, Folksonomy 시스템에서 사용자들이 사용하는 태그들의 빈도수를 토대로 Concept Analyzer의 FFCA를 적용한 결과, 공통 태그를 사용하는 각각의 사용자들이 그 공통 태그를 어느 정도 사용하는지에 대한 정보를 파악 할 수 있었으며, 한 명의 사용자가 북마킹하기 위해 사용한 태그들을 토대로 RFCA를 적용한 결과, 배타적으로 사용된 태그들에 의해 북마킹된 사이트들을 분류 할 수 있었다.

개념분석을 기반으로 하는 향후 연구과제로서는, 개념분석기법을 적용하여 분석된 결과를 토대로, 연관관계(Association rule) 및 함의관계(Implication rule)를 추출하기 위한 제반 연구와 웹상에 존재하는 다양한 데이터에도 개념분석기법을 적용하기 위하여 다양한 웹 마이닝 분석기법들을 접목할 필요가 있다.

참고문헌

[1] Lukasz A. Kurgan and Petr Musilek, "A survey of Knowledge Discovery and Data Mining process models", The Knowledge Engineering Review, Vol. 21:1, pp.1-24, 2006.

[2] Renato Coppi, "A theoretical framework for Data Mining: the "Informational Paradigm", Computation Statistics & Data Analysis, Vol.38, pp.501-515, 2002.

[3] Michael Goebel and Le Gruenwald, "A Survey of Data Mining and Knowledge Discovery Software Tools", ACM SIGKDD, pp. 20-33, 1999.

[4] Margaret H. Dunham, *DATA MINING:Introductory and Advanced Topics*, Prentice hall, 2002.

[5] Sushmita Mitra and Tinku Acharya, *DATA MINING : Multimedia, Soft Computing, and Bioinformatics*, WILEY, 2003.

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

[7] C. Carpineto, G. Romano, Concept Data Analysis: Theory and Applications, Wiley, September, 2004.

[8] B. A. Davey, H.A. Priestley, Introduction to Lattices and Order, Cambridge University Press, 2002.

[9] S. Hwang, H.G. Kim, H.S. Yang, "A FCA-based Ontology Construction for the Design of Class Hierarchy", ICCSA2005, LNCS3481, Springer, 2005.

[10] 황석형, 강유경, 김홍기, 김명기, "FCA를 이용한 임상서식지의 체계화: OO병원의 사례", 대한의료 정보학회지, 제11권, 보완본1호, pp. 53-56, 2005.

[11] Thanh Tho Quan, Siu cheung Hui, and Tru Hoang Cao, "A Fuzzy FCA-based Approach to Conceptual Clustering for Automatic Generation of Concept Hierarchy on Uncertainty Data", International Workshop on Concept Lattice and their Applications, Vol. 110, pp. 1-12, 2004.

[12] Jan Jantzen, *Foundations of Fuzzy Control*, WileyBlackwell; New title edition, 2007.

[13] Thanh Tho Quan, Siu Cheung Hui, and Tru Hoang Cao, "A Fuzzy FCA-based Approach to Conceptual Clustering for Automatic Generation of Concept Hierarchy on Uncertainty Data", International workshop on Concept Lattices and their Applications, Vol. 110, pp. 1-12. 2004.

[14] Zdzislaw Pawlak, *Rough Sets : Theoretical Aspects of Reasoning about Data*, Springer, 1991.

[15] B. Walczak and D. L. Massart, "Tutorial Rough sets theory", Chemometrics and Intelligent Laboratory Systems, Vol. 47, pp. 1-16, 1999.

[16] Jan Komorowski, Lech Polkowsik and Andrzej Skowron, Rough Sets : A Tutorial, in:S.K. Pal, A. New Trend in Decision Making, Springer, Singapore, pp. 1-98, 1999.