제33회 한국정보처리학회 춘계학술발표대회 논문집 제17권 제1호 (2010. 4)

# 퍼지 데이터를 러프 클러스터링하기 위한 도구의 개발

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

# On Developing of a tool for rough clustering fuzzy data

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

요 약

오늘날, World Wide Web의 탄생과 정보통신기술의 비약적인 발전에 의해 매일 방대한 양의 다양한 데이터들이 기하급수적으로 발생되고 있다. 이와 같은 데이터들에는 명확한 경계를 갖는 정보와 더불어서 퍼지정보가 포함되어 있다. 퍼지정보를 포함한 데이터로부터 유용한 정보를 추출하기 위해, 퍼지 데이터 분석 및 러프 데이터 분석에 관한 다양한 연구들이 수행되고 있다. 본 논문에서는, 주어진 퍼지 데이터에 내포된 유용한 정보를 추출하기 위해, 퍼지 집합 이론과 러프 집합 이론을 형식개념분석 기법에 접목하여 새로운 러프 클러스터링 기법을 제안한다. 또한, 본 연구에서 개발한 지원도구와 그 도구를 이용한 실험 결과를 보고한다.

## 1. 서론

통신기술과 컴퓨터의 발달로 지식과 정보의 양은 기하급수적으로 증가하고 있다. 이러한 데이터들은 종종 불완전하거나 불확실하기 때문에 애매한 데이터로부터 유용한 정보를 추출하기가 매우 어려워졌다. 애매모호함을 포함한 대량의 데이터로부터 의미 있는 정보를 추출하기 위한 다양한 데이터 분석기법들이 제안되고 있다[1].

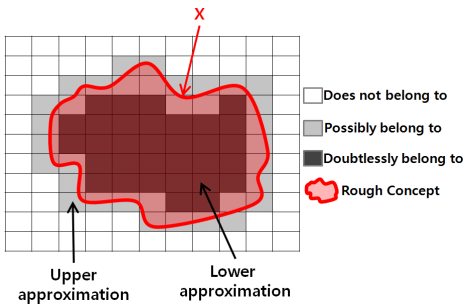
데이터 마이닝(Data Mining)은 데이터에 잠재적으로 내포된 이전에 알려지지 않은 유용한 정보를 추출하는 것이다. 데이터 마이닝은 주어진 데이터로부터 유용한 정보를 추출하고 가시화하는 방법들을 포함한다. 최근, 인간과 기계 사이에 상호 작용을 위한 다양한 방법들 또한 데이터 마이닝의 일부로써 고려된다[1,2]. 분류와 클러스터링은 주어진 데이터로부터 유용한 지식을 추출하기 위해 사용되는 데이터 마이닝의 대표적인 2가지 형태이다. 분류는 데이터를 미리 정의된 클래스에 하나의 아이템으로 분류하는 방법이고, 클러스터링은 다양한 특성을 지닌 데이터들을 사전에 정의된 클래스 없이 유사성에 기반 하여 그룹화하는 방법이다.

특히, 클러스터링을 위한 최신 기법으로서, 다양한 분야에서 활용되고 있는 형식개념분석기법(FCA : Formal Concept Analysis)은 주어진 데이터로부터 정보의 최소단위로써 개념들을 추출하고, 개념들 사이의 관계를 토대로 개념계층구조를 구축하기 위한 수학적 데이터마이닝 기법중의 하나이다. 형식개념분석기법은 대량의 데이터를 체계화 할 수 있으며, 구조화된 정보를 토대로 데이터에 내재된 유용한 정보를 수월하게 추출 할 수 있다[3].

그러나, 형식개념분석기법에서 분석 대상이 되는 데이터는 이진데이터로써 매우 제한적이므로, 퍼지데이터나 러프데이터와 같은 애매한 데이터의 특성을 고려하여 분석하고 처리하기에는 불충분하다. 이와 같이 불명확하고 애매한 데이터를 다루기 위한 방법으로써, 퍼지집합이론[4]과 러프집합이론[5]이 제안되었다.

퍼지집합이론은 애매한 데이터를 정량적으로 처리하기 위한 방법이다. 퍼지집합이론에서 집합의 각 요소는 그 집합에 귀속되어지는 정도를 0부터 1사이의 수로 나타낸 귀속도를 갖는다. 퍼지집합이론은 귀속도에 의해 표현되는 불명확한 정보들을 다룬다. 예를 들어, 초콜릿을 좋아하는 사람들에 대한 데이터를 표현할 때, 보통 집합이론에서는 어떤 사람이 초콜릿을 “좋아한다” 또는 “싫어한다”로만 표현할 수 있지만, 퍼지집합이론에서는 초콜릿을 “80%정도 좋아한다” 또는 “10%정도 좋아한다”와 같이 정량적으로 초콜릿을 좋아하는 정도를 표현할 수 있다. 한편, 러프집합이론은 모호함과 불확실성과 같은 특징을 갖는 데이터를 다루기 위한 또다른 방법으로서 제안되었다. 러프집합이론에서는 애매모호한 데이터를 다루기 위해서, 그림1과 같이, 어떤 집합에 확실하게 분류되는 하한근사(Lower Approximation)와 불확실하게 분류되는 상한근사(Upper Approximation) 사용하여 구별하기 애매모호한 원소들을 하나의 개념으로 포함시켜서 추출함으로써, 주어진 불완전한 데이터를 분류할 수 있다. 최근 데이터베이스의 지식

발견, 패턴 인식, 정보 처리 뿐만 아니라 의료 진단과 의료데이터 분석, 인공지능, 인지과학과 같은 여러 분야에서 응용되고 있으며, 퍼지와 러프집합을 결합한 연구도 활발하게 이루어지고 있다[2].




(그림 1) 상한근사와 하한근사

[6]의 후속연구로서, 본 논문에서는, 주어진 애매모호한 퍼지 데이터에 함축된 의미 있는 정보를 추출하기 위해 퍼지 집합 이론과 러프 집합 이론을 형식개념분석기법에 접목하여 새롭게 제안한 퍼지-러프 클러스터링 기법을 지원하는 도구를 개발하였다. 또한, 퍼지-러프 클러스터링 기법과 본 연구에서 개발된 도구의 유용성과 가능성을 검토하기 위하여, 실세계의 데이터를 대상으로 실험을 수행하고, 그 결과를 보고한다. 본 연구에서 개발한 도구를 사용함으로써, 애매모호한 데이터를 수월하게 분류하고 계층화 할 수 있으며, 이를 토대로 데이터에 내재된 유용한 정보를 추출할 수 있으므로 다양한 분야에서 활용될 수 있다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 퍼지-러프 클러스터링 기법에 대해 설명하고, 3장에서는 본 연구에서 개발한 지원도구 및 실제 데이터에 적용한 실험결과에 대해서 설명한다. 4장에서는 결론과 향후 연구과제에 대해서 설명한다.

## 2. 퍼지-러프 클러스터링 기법

퍼지-러프 클러스터링 기법은 주어진 애매모호한 데이터로부터 숨겨져 있는 지식을 귀속도를 포함한 러프 클러스터링 단위로 추출하여 구조화 할 수 있다. 퍼지-러프 클러스터링 기법에서는 퍼지데이터를 대상으로 Fuzzy context를 사용하여 입력데이터 테이블로 정의한다. 즉, Fuzzy context K := (O, A, R =$\mu(O\times A)$)는 객체들의 집합 O와 속성들의 집합 A, 그리고 O와 A사이의 관계를 나타내는 R로 구성된다. 단, (o, a)∈R은 0과 1사이의 귀속도$\mu(o,a)$를 갖는다. 표1은 4개의 객체들과 4개의 속성들로 구성된 Fuzzy context의 예이다. 관심데이터에 대해서만 퍼지-러프 클러스터링 기법을 적용하기 위해 임계값(Threshold)을 설정할 수 있으며, 임계값 T를 기준으로 입력된 데이터를 필터링하여 간략화 된 Fuzzy context를 도출해 낼 수 있다. 표2는 표1에서 임계값 T를 0.5로 설정하여 필터링된 Fuzzy context이다.

<표 1> Fuzzy context

| Objects \ Attributes | a | b | c | d |
|---|---|---|---|---|
| o1 | 0.0 | 0.5 | 0.1 | 0.7 |
| o2 | 0.0 | 0.4 | 0.9 | 0.0 |
| o3 | 0.3 | 0.0 | 0.5 | 0.0 |
| o4 | 0.6 | 0.8 | 0.2 | 0.2 |

<표 2> 임계값 T=0.5에 의해 필터링 된 Fuzzy context

| Objects \ Attributes | a | b | c | d |
|---|---|---|---|---|
| o1 | - | 0.5 | - | 0.7 |
| o2 | - | - | 0.9 | - |
| o3 | - | - | 0.5 | - |
| o4 | 0.6 | 0.8 | - | - |

주어진 퍼지 데이터를 러프 클러스터링 하기 위해 러프집합이론에서 제공하는 상한근사와 하한근사 함수를 퍼지-러프 클러스터링 기법에 적합하도록 다음과 같이 재정의하였다.

[정의1] 임의의 Fuzzy context K := (O, A, R =$\mu(O\times A)$) 에 대해서, X⊆O, Y⊆A 일 때, 하한근사(LA)와 상한근사(UA)는 다음과 같이 정의한다.

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

단, OS(a):={o∈O|(o, a)∈R}, AS(o):={a∈A|(o, a)∈R}.

LA(X)는 어떤 집합 X에 확실하게 포함되는 원소들이 갖는 배타적 속성의 집합, 즉, X에 완전하게 포함되는 객체들이 배타적으로 갖는 속성들의 집합이다. UA(X)는 어떤 집합X에 불확실하게(또는 불완전하게) 속하는 원소들이 갖는 총체적인 속성의 집합, 즉, X에 총체적으로 속하는 객체들이 갖는 속성들의 집합이다. 위와 마찬가지로, 어떤 집합 Y에 대해서, LA함수와 UA함수를 적용하면 각각 집합 Y에 확실하게 포함되는 속성들을 배타적으로 갖는 객체집합과 Y에 불확실하게 속하는 속성들을 총체적으로 갖는 객체집합을 구할 수 있다. 예를 들어, 표2에 대해서, X = {o1, o2, o3}, Y = {a, c} 일 때, LA(X) = {c, d}이고, UA(X) = {b, c, d}이며, LA(Y) = {o2, o3}이고, UA(Y)={o2, o3, o4}이다.

LA와 UA함수를 사용하여 주어진 Fuzzy context로부터 러프개념을 추출할 수 있다.

[정의2] 임의의 Fuzzy context K := (O, A, R =$\mu(O\times A)$) 에 대해서, X⊆O, Y⊆A 일 때, X = UA(Y) ∧ Y = LA(X)를 만족하는 (X, Y, Z)를 러프개념라고 한다. 즉, (X, Y) = (UA(LA(X)), LA(X)), Z = {(x, y, r)| x∈X ∧ y∈ Y ∧ r = $\mu(x,y)$}.

퍼지-러프 클러스터링 기법에서는, 하나의 러프개념(X, Y, Z)은 3개의 요소로 구성된다. Y에 불확실하게 속하는 속성들을 총체적으로 갖는 객체들의 집합 X와 X에 완전하게 포함되는 객체들이 배타적으로 갖는 속성들의 집합 Y, 그리고 Z는 X와 Y사이의 관계를 나타내는 집합이다.

예를 들어, 표2에 대해서, X = {o1, o2, o3} 일 때, LA(X) = {c, d}이고, UA(LA(X)) = UA({c, d}) = {o1, o2, o3}, Z = {(o1, d, 0.7), (o2, c, 0.9), (o3, c, 0.5)}이다. 즉, X = UA(Y) ∧ Y = LA(X)이므로, ({o1, o2, o3}, {c, d}, {(o1, d, 0.7), (o2, c, 0.9), (o3, c, 0.5)})는 러프개념이다. 이와 같은 방법으로 표2로부터 8개의 러프개념들이 표3과 같이 추출되었다.

<표 3> 표2로부터 추출된 러프개념들

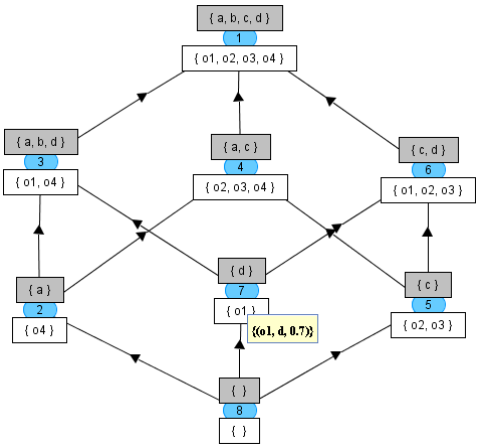
| Rough Concepts | Extension | Intension | Relations |
|---|---|---|---|
| $C_1$ | {o1, o2, o3, o4} | {a, b, c, d} | {(o1, b, 0.5), (o1, d, 0.7), (o2, c, 0.9), (o3, c, 0.5), (o4, a, 0.6), (o4, b, 0.8)} |
| $C_2$ | {o4} | {a} | {(o4, a, 0.6)} |
| $C_3$ | {o1, o4} | {a, b, d} | {(o1, b, 0.5), (o1, d, 0.7), (o4, a, 0.6), (o4, b, 0.8)} |
| $C_4$ | {o2, o3, o4} | {a, c} | {(o2, c, 0.9), (o3, c, 0.5), (o4, a, 0.6)} |
| $C_5$ | {o2, o3} | {c} | {(o2, c, 0.9), (o3, c, 0.5)} |
| $C_6$ | {o1, o2, o3} | {c, d} | {(o1, d, 0.7), (o2, c, 0.9), (o3, c, 0.5)} |
| $C_7$ | {o1} | {d} | {(o1, d, 0.7)} |
| $C_8$ | {} | {} | {} |

정의2에서 추출된 러프개념들 사이에는 "sub-concept of"관계가 존재한다.

[정의3]임의의 러프개념 $C_1 = (X_1, Y_1, Z_1)$, $C_2 = (X_2, Y_2, Z_2)$에 대하여, "sub-concept of"관계 $(X_1, Y_1, Z_1) \lesssim (X_2, Y_2, Z_2)$는 다음과 같이 정의 된다.

$(X_1, Y_1, Z_1) \lesssim (X_2, Y_2, Z_2) \Leftrightarrow X_1 \subseteq X_2 (\Leftrightarrow Y_1 \subseteq Y_2, \Leftrightarrow Z_1 \subseteq Z_2)$.

"sub-concept of"관계는 러프개념을 구성하는 객체들과 속성들 사이의 부분집합 관계를 나타낸다. 예를 들어, 두 러프개념 $C_5$ = ({o2, o3}, {c}, {(o2, c, 0.9), (o3, c, 0.5)})와 $C_6$ = ({o1, o2, o3}, {c, d}, {(o1, d, 0.7), (o2, c, 0.9), (o3, c, 0.5)})에 대해서, {o2, o3} ⊆ {o1, o2, o3} (⇔ {c} ⊆ {c, d}, ⇔ {(o2, c, 0.9), (o3, c, 0.5)} ⊆ {(o1, d, 0.7), (o2, c, 0.9), (o3, c, 0.5)})이므로, $C_5$는 $C_6$의 하위개념이다.

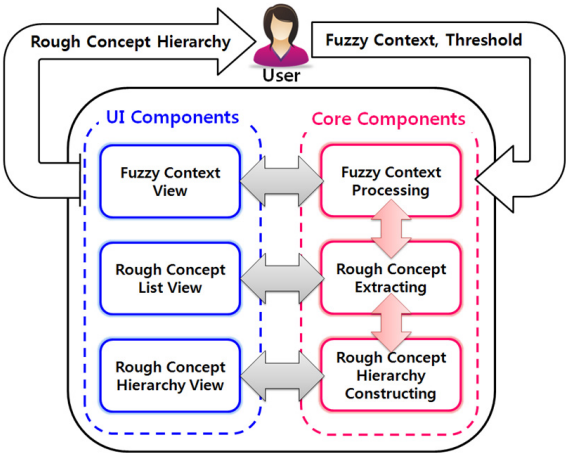



(그림 2) 표2에 대한 Rough Concept Hierarchy

위의 정의들을 토대로 구축된 Rough Concept Hierarchy는 그림2와 같다. Rough Concept Hierarchy에서, 각 노드는 러프개념을 나타내고 노드들 사이에 연결된 선은 "sub-concept of"관계를 나타낸다. 노드 위의 레이블은 러프개념의 Intension 정보를 나타내고, 노드 아래의 레이블은 Extension 정보를 나타낸다. 노드를 선택하면 선택된 노드의 Extension과 Intension 사이의 관계 정보를 확인할 수 있다. 그림2에서, 최상위 개념 $C_1$은 주어진 Fuzzy context의 객체와 속성의 전체집합을 나타낸다. $C_1$은 3개의 하위개념 $C_3$, $C_4$, $C_6$로 분류되며, 3개의 러프개념들은 각각 2개의 러프개념들로 분류된다.

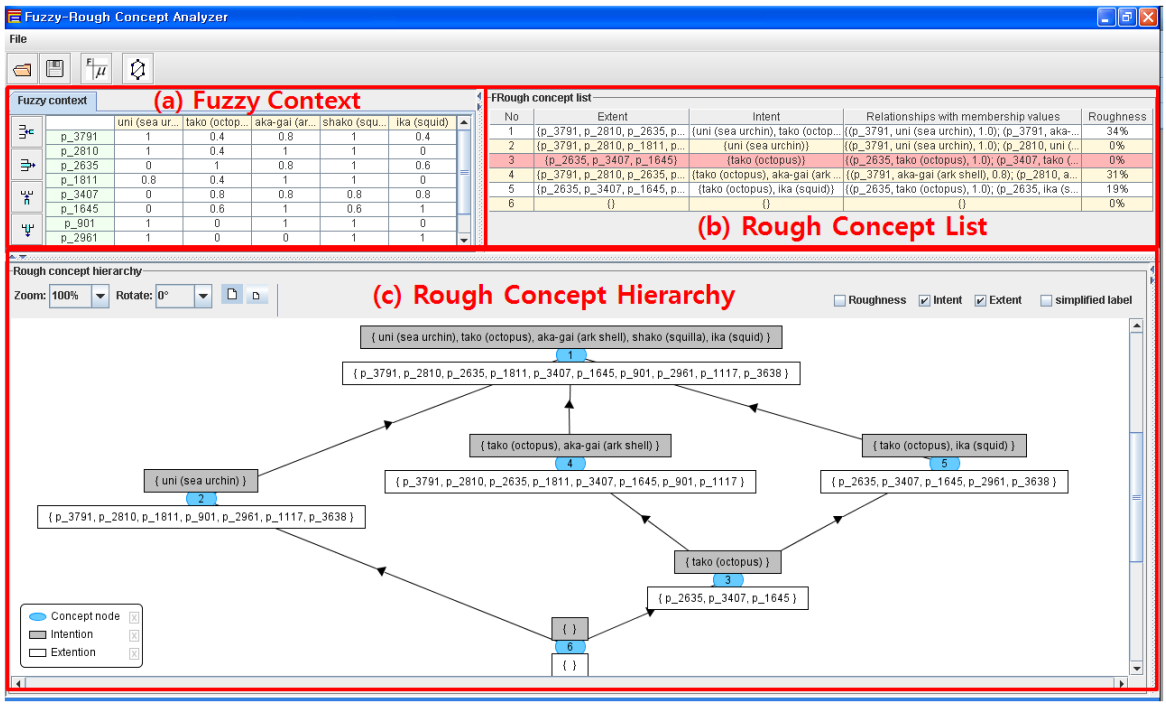
## 3. Fuzzy-Rough Concept Analyzer

본 장에서는, 앞 절의 제반 정의들을 토대로, 퍼지-러프 클러스터링 기법을 지원하기 위해 개발된 FuRoCA(Fuzzy-Rough Concept Analyzer)를 소개한다. 또한, 실제 데이터에 FuRoCA를 사용하여 퍼지-러프 클러스터링 기법을 적용한 실험에 대해서 설명한다.




(그림 3) FuRoCA의 아키텍처

FuRoCA의 전체적인 아키텍처는 그림3과 같이 사용자 편의를 위한 UI 컴포넌트와 내부 데이터 처리를 위한 코어 컴포넌트로 구성되어 있다. 코어 컴포넌트는 주어진 데이터를 Fuzzy context로 변환하고 임계값에 의해 필터링된 Fuzzy context를 도출하기 위한 Fuzzy Context Processing 컴포넌트와, Fuzzy context로부터 러프개념들을 추출하기 위한 Rough Concept Extracting 컴포넌트, 그리고, 추출된 러프개념들과 그들 사이의 관계를 파악하여 Hierarchy를 구축하기 위한 Rough Concept Hierarchy Constructing 컴포넌트로 구성되어있다. 한편, UI 컴포넌트는 다음과 같은 3개의 서브 뷰로 구성되어있다. Fuzzy Context View는 그림4의 (a)와 같이 입력된 데이터를 Fuzzy context형태로 보여줄 뿐만 아니라 새로운 context를 생성하거나 편집하기 위한 기능도 제공한다. Rough Concept List View는 Rough Concept Extracting 컴포넌트로부터 추출된 모든 러프개념들의 정보를 그림4의 (b)와 같이 테이블 형태로 보여준다. Rough Concept




(그림 4) FuRoCA의 실행 화면

Hierarchy View는 Rough Concept Hierarchy Constructing에 의해 구축된 Hierarchy를 그래픽하게 가시화한다.

FuRoCA의 유용성과 가능성을 검토하기 위해, 5000명의 응답자가 100개의 스시에 대한 선호도를 조사한 실제 설문조사 데이터를 대상으로 실험을 실시하였다(http://www.kamishima.net/sushi). 그림4의 (a)는 5000명의 설문응답자들이 가장 선호하는 5개의 스시를 가장 좋아하는 10명의 응답자에 대한 데이터로써, 응답자가 각각의 스시를 어느 정도 좋아하는지에 관한 선호도 정보를 표현한 Fuzzy context이다. 예를 들어, 응답자 p_3791은 aka-gai를 80%(선호도 : 0.8)정도 선호한다. 선호도가 낮은 데이터를 제거하기 위해 임계값 T를 0.5로 설정하여 필터링 한 후 러프개념들을 추출하여 Rough Concept Hierarchy를 구축하였다(그림4의 (c)참조). 예를 들어, $C_5$는 5명의 응답자 p_2635, p_3407, p_1645, p_2961, p_3638가 다른 응답자들과는 배타적으로 선호하는 스시가 tako와 ika임을 나타내는 러프개념이다. 또한, $C_3$는 tako 스시를 배타적으로 선호하는 사람이 응답자중 3명(p_2635, p_3407, p_1645)임을 나타낸다.

## 4. 결론

본 논문에서는 퍼지정보가 포함된 데이터로부터 수월하게 유용한 정보를 추출하기 위해, 퍼지-러프 클러스터링 기법을 제안하고, 이를 지원하기 위한 도구를 개발하였다. 또한, 본 연구결과의 가능성을 검토하기위해 실제 퍼지정보를 포함한 스시 선호도 설문데이터를 대상으로 퍼지-러프 클러스터링 기법을 적용하는 실험을 수행하여 Rough Concept Hierarchy를 구축하였다. 실험 결과를 토대로, 다른 설문응답자들과는 구별되게 배타적으로 선호하는 스시를 추출해 낼 수 있을 뿐만 아니라 어느 정도 선호하는지에 대한 정량적인 선호도 정보도 확인할 수 있었다. 배타적으로 선호하는 스시에 의해서 응답자들을 클러스터링함으로서, 응답자들의 스시 선호도에 관한 배타적 성향을 파악해 볼 수 있다.

향후 연구과제로서, 의료데이터 또는 웹 데이터와 같이 애매모호함을 포함하는 대량의 데이터를 대상으로 퍼지-러프 클러스터링 기법을 적용하여 분석하기 위해, FuRoCA의 성능을 향상할 필요가 있다. 또한, 특정 조건에 맞는 러프개념을 추출하기 위한 쿼리기능 뿐만 아니라 러프개념들 사이에 존재하는 연관규칙을 추출하기 위한 기능도 추가할 계획이다.

## 참고문헌

[1] Frawley, W.J., Shapiro, G.P. and Matheus, C.J. Knowledge Discovery in Databases: An Overview. AI Magazine, 213-228, 1992.

[2] S. D. Jitender, V. V. Raghavan, A. Sarkar and H. Sever, “Data Mining Trends in Research and Development”, Rough Sets and Data Mining analysis of imprecise data, pp. 9-45, 1997.

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

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

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

[6] Yu-Kyung Kang, Suk-Hyung Hwang, and Hae-Sool Yang, "A FCA-based Classification of uncertainty data using Rough Clustering", ICCI09, pp.270-274, 2009.