퍼지 데이터로부터 연관 규칙을 추출하기 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 association rule extracting from 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, 대량의 데이터를 수집, 저장 및 관리하는 데이터베이스 기술의 진보를 기반으로, 의료, 과학, 교육, 비즈니스 등 다양한 분야에서 발생되는 대규모 데이터를 축적하게 되었다. 다양한 분야에서 축적된 대량의 데이터에 내재된 유용한 정보를 수월하게 추출하여 분석하기 위해 널리 사용되고 있는 형식개념분석기법은, 주어진 데이터로부터 정보의 최소단위로써 개념들을 추출하고, 개념들 사이의 관계를 토대로 개념계층구조를 구축하기 위한 정형화된 데이터마이닝 기법을 제공하고 있다. 본 논문에서는, 주어진 퍼지 데이터에 잠재된 유용한 정보를 추출하기 위해, 퍼지 집합 이론을 형식개념분석기법에 접목한 퍼지개념분석기법과 이를 지원하기 위해 본 연구에서 개발된 FFCA-Wizard를 소개한다. 또한, FFCA-Wizard를 사용하여 실세계 데이터를 대상으로 퍼지개념분석을 실시한 실험 결과를 보고한다.

1. 서론

데이터 생성, 수집과 저장 기술의 급속한 발전으로 인하여 상거래와 다양한 과학 분야에서 방대한 양의 데이터 집합이 생성되고 기관들은 이러한 대규모 데이터를 축적하게 되었다. 축적된 데이터의 패턴 및 성향을 분석하여 도메인의 현재 상황 및 경향 파악, 향후 전망에 활용하기 위해 다양한 데이터분석 기법들이 제안되었다[1].

데이터 마이닝은 분류 기법(Classification), 군집화 기법(Clustering), 요약 기법(Summarizing), 순차적 패턴 분석(Sequential pattern discovery), 연관 규칙 추출(Association rule discovery) 등을 아우르는, 데이터로부터 유용한 정보를 추출하는 기술을 통틀어 일컫는다. 군집화 기법 중 최근 다양한 분야에서 널리 활용되고 있는 형식개념분석기법(FCA : Formal Concept Analysis)은 주어진 데이터로부터 객체(Object)와 속성(Attribute)들을 추출하고, 이들 사이의 포함관계를 파악하여, 개념(Concept)을 생성하고, 추출된 개념들 사이에 상•하위 관계를 파악하여 개념계층구조(Concept hierarchy)를 구축하기 위한 수학적 데이터마이닝 기법중의 하나이다. 형식개념분석기법은 대량의 데이터를 체계화 할 수 있으며, 구조화된 정보를 토대로 데이터에 내재된 유용한 정보를 수월하게 추출할 수 있다[2].

한편, 퍼지 집합 이론(Fuzzy Set Theory)[3]은 '네' 또는 '아니오' 등 이분법으로는 나타내기 힘든 인간의 언어와 사고에 관련된 "애매함"과 "모호성"을 정량적으로 표현해 내는데 쓰인다. 퍼지 집합 이론에서 집합의 각 원소는 그 집합에 귀속되어지는 정도를 0부터 1사이의 수로 나타낸 귀속도(歸屬度, membership degree)를 갖는다. 예를 들어, 초밥을 좋아하는 사람들에 대한 데이터를 표현할 때, 보통 집합이론에서는 어떤 사람이 초밥을 "좋아한다" 또는 "싫어한다"로만 표현할 수 있지만, 퍼지 집합 이론에서는 초밥을 "70%정도 좋아한다" 또는 "20%정도 좋아한다"와 같이 정량적으로 초밥을 좋아하는 정도를 표현할 수 있다.

본 논문에서는, 주어진 애매모호한 퍼지 데이터에 잠재적으로 내포된 유용한 정보를 추출하고 데이터들 사이의 연관 관계를 파악하기 위해 퍼지개념분석[4]과 연관 규칙[5]을 소개하고, 이를 지원하는 도구를 개발하였다. 또한, 본 연구에서 개발된 도구의 유용성과 가능성을 검토하기 위하여, 실세계의 데이터를 대상으로 실험을 수행하고, 그 결과를 보고한다. 본 연구에서 개발한 도구를 사용함으로써, 애매모호한 퍼지데이터를 수월하게 분류하고 계층화할 수 있으며, 이를 토대로 데이터에 내재된 유용한 정보를 추출할 수 있다. 뿐만 아니라 데이터들 사이에 내제된 연관 규칙들을 추출함으로써, 도메인의 경향 파악 및 향후 전망을 예측하는데 사용될 수 있으므로 다양한 분야에서 활용될 수 있다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 퍼지개념분석기법과 연관 규칙에 대해 설명하고, 3장에서는 본 연구에서 개발한 지원도구 및 실제 데이터에 적용한 실험 결과에 대해서 설명한다. 4장에서는 결론과 향후 연구과제에 대해서 설명한다.

2. 퍼지개념분석기법과 연관 규칙

본 장에서는, 주어진 애매모호한 퍼지 데이터에 함축된 의미 있는 정보를 추출하고 추출된 퍼지 개념들 사이에 연관관계를 파악하기 위한 퍼지개념분석기법(Fuzzy Formal Concept Analysis)과 연관 규칙(Association rule)을 설명한다.

2.1 퍼지개념분석기법

퍼지개념분석기법[4]에서는 주어진 도메인의 퍼지데이터(어떤 객체가 갖는 속성과 해당객체의 속성에 대한 귀속도)를 대상으로 Fuzzy context를 사용하여 입력데이터 테이블로 정의한다.

[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 귀속도 μ(g, m)를 갖는다.■

표1은 4개의 객체들과 4개의 속성들로 구성된 Fuzzy context의 예이다. 관심데이터에 대해서만 퍼지개념분석기법을 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 0.6으로 설정하여 필터링된 Fuzzy context이다.

<표 1> Fuzzy context

Objects \ Attributes	a	b	c	d
o1	1.0	0.0	0.4	0.0
o2	0.5	1.0	0.6	1.0
o3	0.9	0.6	0.1	0.9
o4	0.3	1.0	0.8	0.7

<표 2> 임계값 T = 0.6에 의해 필터링 된 Fuzzy context

Objects \ Attributes	a	b	c	d
o1	1.0	-	-	-
o2	-	1.0	0.6	1.0
o3	0.9	0.6	-	0.9
o4	-	1.0	0.8	0.7

[정의 2] 임의의 Fuzzy context K := (G, M, I=φ(G×M))와 임계값 T에 대하여, A⊆G, B⊆M일 때, FE(B) = A ∧ B = FI(A)를 만족하는 (φ(A), B)를 퍼지개념이라고 한다. 단, FI(A) = {m∈M | ∀g∈A : μ(g, m)≥T}, FE(B) = {g∈G | ∀m∈B : μ(g, m)≥T}. 또한, 각 객체 g∈φ(A)에 대해서 $\mu_g = \min_{m \in B} \mu(g, m)$ 이다.■

즉, μ_g는 객체 g와 속성 m사이의 귀속도 중 최소값을 나타낸다. 만약, B = {} 라면 모든 g∈A에 대한 μ_g = 1 이다. 표2로부터 추출된 모든 퍼지개념들은 표3과 같다 (T=0.6). 예를 들어, 표2에 대해서, A = {o3}이고, B = {a, b, d}일 때, FI(o3) = {a, b, d}이고, FE({a, b, d}) = {o3} 이므로 ({o3}, {a, b, d})는 퍼지개념이다. 단, μ(o3, a) = 0.9이고, μ(o3, b) = 0.6이고, μ(o3, d) = 0.9이므로, μ_{o3} = 0.6이다. 따라서, ({o3(0.6)}, {a, b, d})와 같이 표현된다.

Fuzzy context K로부터 추출된 모든 퍼지개념들 사이에는 다음과 같은 상·하위개념관계(Super-sub concept relation)가 존재한다.

<표 3> 표2에서 추출된 퍼지개념들

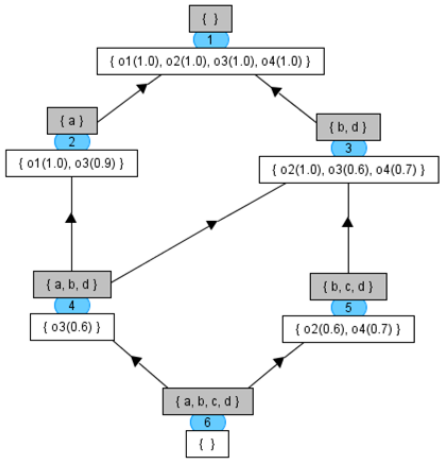
ID	Extension	Intension
C1	{o1(1.0), o2(1.0), o3(1.0), o4(1.0)}	{ }
C2	{o1(1.0), o3(0.9)}	{a}
C3	{o2(1.0), o3(0.6), o4(0.7)}	{b, d}
C4	{o3(0.6)}	{a, b, d}
C5	{o2(0.6), o4(0.7)}	{b, c, d}
C6	{ }	{a, b, c, d}

[정의 3] 임의의 퍼지개념 C1=(φ(A1), B1), C2=(φ(A2), B2)에 대하여, 상·하위개념관계 (φ(A1), B1)≤(φ(A2), B2)는 다음과 같이 정의 된다.

(φ(A1),B1)≤(φ(A2), B2)⇔φ(A1)⊆φ(A2)(⇔B1⊇B2).■

퍼지개념 C2 = ({o2(1.0), o3(0.6), o4(0.7)}, {b, d})와 C4 = ({o3(0.6)}, {a, b, d})에 대해서, {o3(0.6)}⊆{o2(1.0), o3(0.6), o4(0.7)}(⇔{a, b, d}⊇{b, d})이므로, C3는 C4의 상위개념(Super Concept)이며, C4 ≤ C3과 같이 표현한다.

임의의 Fuzzy context K에 대한 퍼지개념격자 B(K)는 K에 대한 임계값 T와 K로부터 추출된 모든 퍼지개념들과 그들 사이의 상·하위개념관계들에 의해 그림1과 같이 표현된다.



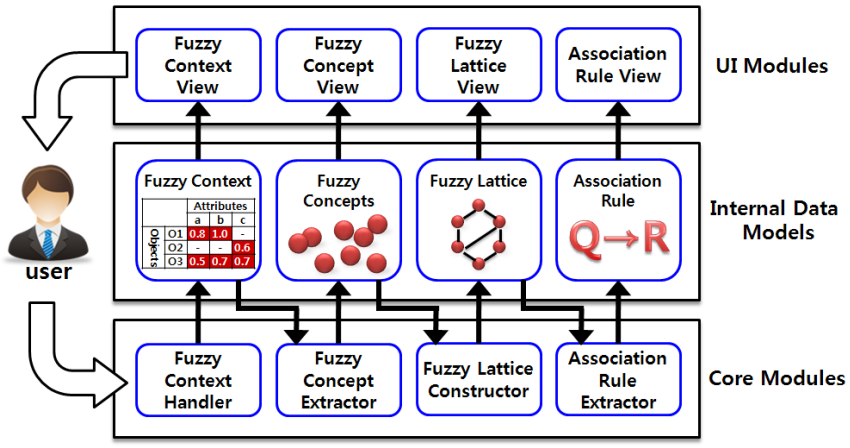
(그림 1) 표2에 대한 퍼지개념격자

퍼지개념격자는 정점들과 링크들로 구성되며, 각 정점은 퍼지개념들에 대응하고, 각 정점에는 2종류에 레이블(Extension과 Intension)이 각각 하단과 상단에 표시된다. 각 퍼지개념들 사이를 연결하는 링크들은 개념들 사이의 "상·하위개념관계"를 나타낸다.

2.2 연관 규칙

연관 규칙(Association rule) 추출 기법에 관한 기본 이론은, 대형 슈퍼마켓에 비치된 전산화된 계산대(Point of sale system)로부터 수집된 대용량의 상품 구매 정보를 통해, 상품 간의 동시 구매 관계를 분석하여 이를 마케팅 전략 등에 활용하고자 1993년 Agrawal 등에 의해 제안되었다[5]. 추출된 연관 규칙 A→B는 "상품 집합 A를 구입하는 고객은, 상품 집합 B를 구입하는 경향이 있다."라고 해석된다. 퍼지개념분석기법 기반의 연관 규칙 추출을 위해, 연관 규칙을 다음과 같이 재정의 하였다.

[정의 4] 주어진 Fuzzy context K := (G, M, I=φ(G×M))



(그림 2) FFCA-Wizard의 아키텍처

의 임의의 두 속성 Q, R⊆M이, |FE(Q∪R)|/|G|≥minsup와 |FE(Q∪R)|/|FE(Q)|≥minconf를 만족하는 경우, Q는 R과 연관된다 라고 하며, $Q \rightarrow R_{minsup,minconf}$로 표기한다. 단, minsup, minconf∈[0, 1].■

정의 4에서 언급된 |FE(Q∪R)|/|G|과 |FE(Q∪R)|/|FE(Q)|는, 각각 연관 규칙 Q→R의 지지도(support)와 확신도(confidence)라고 부른다. 특히 minsup(Minimum Support : 최소 지지도)와 minconf(Minimum Confidence : 최소 확신도)는 집합 Q가 R과 갖는 연관관계를 바라보는 분석가의 주관적인 경계값(Threshold)이다. 먼저 minsup는 Q와 R사이에 존재하는 관계가 전체 객체들 중 적어도 얼마나 되는 객체 사이에서 성립할 경우에 이 두 집합 간의 연관관계를 인정할 것인가에 대한 경계값이다. 그리고 minconf는 Q라는 속성을 갖는 객체 중 적어도 얼마나 되는 객체가 R이라는 속성 역시 가질 때, Q는 R과 연관관계에 있다고 판단할 것인지에 대한 경계값이다.

예를 들어, 표2의 Fuzzy context에서, Q={b, d}, R={c}이라 하고, minsup = 0.5, minconf = 0.5이라 하면, |FE({b, c, d})|/|G| = |{o2,o4}|/|4| = 2/4 =0.5≥minsup, |FE({b, c, d})|/|FE({b, d})| = |{o2, o4}|/|{o2, o3, o4}| = 2/3 ≒ 0.667≥minconf 이므로, Q→R 즉, 50%이상의 지지도와 50%이상의 확신도에서 {b, d}→{c}는 성립한다고 할 수 있다. 이와 같이, 연관규칙 {b, d}→{c}는 '속성 b와 c를 갖는 객체 중 66% 이상이 속성 c를 갖는 경향을 보이며, 이러한 현상은 전체 객체 중 50% 이상의 객체로부터 관찰 된다.'라고 해석할 수 있다. 연관 규칙 중 minsup = 0, minconf = 1인 경우 추출되는 연관 규칙을 함의관계(Implication)이라고 한다. 함의관계는 '속성 집합 Q를 갖는 객체들은 반드시 속성 집합 R도 갖는다.'라는 보다 엄격한 관계를 나타낸다.

3. 지원도구 및 실험

본 장에서는, 앞 절의 제반 정의들을 토대로, 퍼지개념분석기법을 지원하기 위해 개발된 FFCA-Wizard(Fuzzy Formal Concept Wizard)를 소개한다. 또한, 실제 데이터에 FFCA-Wizard를 사용하여 퍼지개념분석기법을 적용하여 연관 규칙을 추출한 실험에 대해서 설명한다.

FFCA-Wizard의 전체 아키텍처는 그림2와 같이 3계층 구조로 구성되어 있으며, 각 계층의 모듈들에 대한 설명은 다음과 같다.

(1) Core Modules

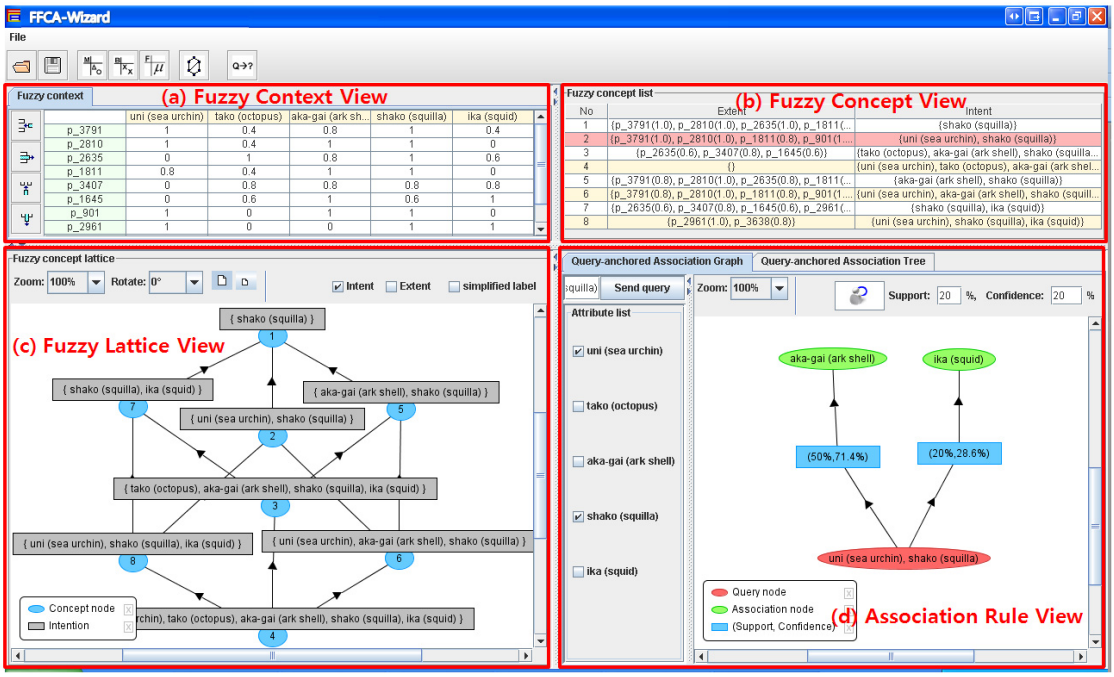
- Fuzzy Context Handler: 입력된 퍼지 데이터를 테이블 형태인 Fuzzy context로 표현하는 모듈로서, 분석하고자 하는 관점에 따라서 임계값을 설정하여 데이터를 필터링 할 수 있다.
- Fuzzy Concept Extractor : 주어진 fuzzy context로부터 퍼지 개념 및 그들 사이의 관계를 추출하는 모듈이다.
- Fuzzy Lattice Constructor : 추출된 퍼지 개념들과 그들 상이의 상•하위 관계를 파악하여 퍼지개념격자를 구축하는 모듈이다.
- Association Rule Extractor : 주어진 Fuzzy context와 Fuzzy lattice를 토대로, 속성들 사이에 최소 지지도와 최소 확신도를 만족하는 연관 규칙을 추출하는 모듈이다.

(2) Internal Data Models

- Fuzzy Context : 주어진 도메인으로부터 객체와 속성들을 추출하여 cross table 형태로 표현하고 객체와 속성 사이의 관계를 해당 셀에 [0, 1]사이의 귀속도를 표시한다.
- Fuzzy Concepts : 객체와 속성, 귀속도로 구성되며, 공통 속성을 갖는 객체들의 클러스터이고, 퍼지 개념에 표현되는 귀속도는 객체와 속성 사이의 귀속도 중 최소값을 나타낸다.
- Fuzzy Lattice : 추출된 퍼지 개념들 사이의 상•하위 관계를 파악하여 구축된 개념격자이다.
- Association Rule : 유용성과 확실성을 반영하여 속성들 사이에 유용한 연관성과 상관관계를 찾아낸다.

(3) UI Modules

- Fuzzy Context View : 입력된 퍼지 데이터를 Fuzzy Context Handler를 적용하여 사용자에게 Fuzzy Context 형태로 나타낸다(그림3의 (a)). 또한, Fuzzy context 생성 및 편집 기능도 제공한다.
- Fuzzy Concept View : Fuzzy Concept Extractor에서 추출된 모든 개념들에 대한 정보를 사용자에게 표3과 같은 형태로 보여준다(그림3의 (b)).
- Fuzzy Lattice View : Fuzzy Lattice Constructor에 의해 생성된 퍼지개념격자를 가시화한다(그림3의 (c)).
- Association Rule View : 추출된 Association Rule들을 그래프 형태로 지지도와 확신도의 정보를 포함하여 가시



(그림 3) FFCA-Wizard 실행화면

화한다(그림3의 (d)).

FFCA-Wizard의 유용성과 가능성을 검토하기 위해, 5000명의 응답자가 100개의 스시에 대한 선호도를 조사한 실제 설문조사 데이터(http://www.kamishima.net/sushi)를 대상으로 실험을 실시하였다. 그림3의 (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_2635는 ika를 60%(선호도:0.6)정도 선호한다. 선호도가 낮은 데이터를 제거하기 위해 임계값 T를 0.5로 설정하여 필터링 한 후 8개의 퍼지개념들을 추출하여 Fuzzy Lattice를 구축하였다(그림3의 (c)참조). 예를 들어, C6은 3개의 스시 uni, aka-gai, shako를 5명의 응답자 중 p_3791과 p_1811이 80%이상 선호하고, p_2810, p_901, p_117은 100% 선호함을 나타내는 퍼지개념이다. 그림3의 (d)는 최소 지지도 20%와 최소 확신도 20%를 만족하는 연관규칙을 추출하여 그래프로 표현한 것이다. 스시 uni와 shako를 선호하는 응답자들 중 71.4% 이상이 스시 aka-gai를 선호하는 경향을 보이며, 이러한 현상은 전체 응답자 중 50% 이상의 응답자로부터 관찰된다. 또한, 스시 uni와 shako를 선호하는 응답자들 중 28.6% 이상이 스시 ika를 선호하는 경향을 보이며, 이러한 현상은 전체 응답자 중 20% 이상의 응답자로부터 관찰된다.

4. 결론

본 논문에서는 퍼지정보가 포함된 데이터로부터 수월하게 유용한 정보를 추출하고 그들 사이의 연관 관계를 파악하기 위해, 퍼지개념분석기법과 연관규칙을 소개하고, 이를 지원하기 위한 도구를 개발하였다. 또한, 본 연구결과의 가능성을 검토하기위해 실제 퍼지정보를 포함한 스시 선호도 설문데이터를 대상으로 퍼지개념분석기법을 적용하는 실험을 수행하여 Fuzzy Lattice를 구축하였으며, 연관 규칙들도 추출하였다.

실험 결과를 토대로, 응답자들 중 공통으로 선호하는 스시를 추출해 낼 수 있을 뿐만 아니라 어느 정도 선호하는지에 대한 정량적인 선호도 정보를 확인할 수 있었다. 또한, 특정 스시들을 선호하는 응답자들이 최소 지지도와 최소 확신도를 만족하는 범위 내에서 선호하는 연관 관계가 있는 다른 스시들을 추출해 낼 수 있었다.

본 연구에서 개발된 FFCA-Wizard를 사용함으로써, 주어진 퍼지 데이터를 개념적으로 구조화 할 수 있고, 데이터의 경향을 파악하여 향후를 예측할 수 있다. 또한, 추출된 연관규칙들을 토대로 비슷한 성향의 사람에게 알맞은 스시를 추천해 줄 수 있으며, 다양한 분야의 추천 시스템에서 활용할 수 있다. 향후 연구과제로서, 논리연산을 조합하여 다양한 연관규칙을 추출하기 위한 기능을 추가할 예정이며, 웹 데이터와 같이 애매모호함을 포함하는 대량의 데이터를 처리하기 위해 FFCA-Wizard의 성능을 향상할 계획이다.

참고문헌

[1] T. Pang-Ning, Steinbach, Michael, Kumar, Vipin, Data Mining, Addison-Wesley, 2005.

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

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

[4] T. T. Quan, S. C. Hui and T. H. Cao, "A Fuzzy FCA-based Approach to Conceptual Clustering for Automatic Generation of Concept Hierarchy on Uncertainty Data", in Proc. CLA, pp. 1-12, 2004.

[5] A. Rakesh, S. Ramakrishnan, "Fast algorithms for mining association rules", The VLDB confersence, pp.487-499, 1994.