형식개념분석기법을 이용한 폭소노미 데이터 마이닝

강유경*, 황석형*, 양해술**
*선문대학교 컴퓨터공학부, **호서대학교 벤처전문대학원
e-mail: {aquamint99, shwang}@sunmoon.ac.kr, hsyang@office.hoseo.ac.kr

Folksonomy Data Mining using Formal Concept Analysis

Yu-Kyung Kang*, Suk-Hyung Hwang*, Hae-Sool Yang**
*Dept. of Computer Science & Engineering, SunMoon University,
**Graduate School of Venture, Hoseo University

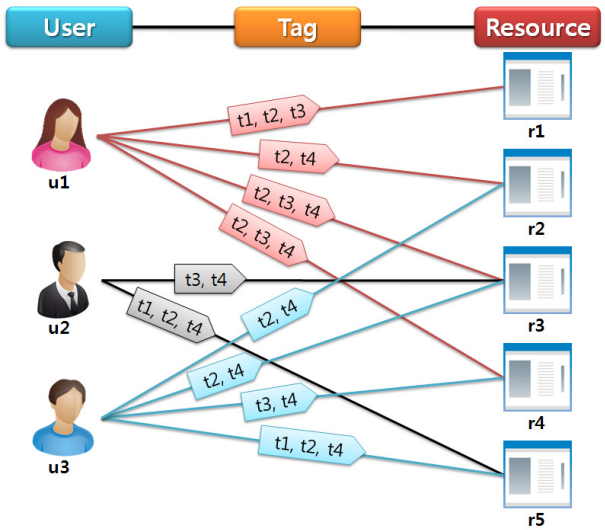
요 약

웹 2.0시대의 대표적인 특징인 폭소노미(folksonomy)는 웹에 존재하는 리소스에 대해 구성원이 자유롭게 선택한 태그(tag)를 붙여서 정보를 체계화하는 새로운 분류 체계이다. 폭소노미를 기반으로하는 웹 애플리케이션 시스템에는 WWW를 이용하는 전 세계의 수많은 사용자들의 다양한 데이터가 축적되어 있으며, 이러한 웹 데이터는 계속적으로 증가·확장·변화하고 있다. 본 논문에서는, 방대한 양의 폭소노미 데이터로부터 유용한 정보를 추출하기 위해 형식개념분석기법을 기반으로, 사용자, 태그, 리소스들 사이의 3항관계를 고려한 폭소노미 데이터 마이닝 기법을 제안하고, 본 연구에서 제안한 기법을 BibSonomy의 데이터에 적용하여 분석한 실험 결과를 보고한다.

1. 서론

오늘날 World Wide Web의 발달과 더불어 웹2.0 시대가 도래하면서 하루에도 수많은 커뮤니티들과 다양한 콘텐츠들이 생성되고 있다. 웹 2.0시대의 대표적인 특징인 폭소노미(folksonomy)는 웹에 존재하는 리소스에 대해 구성원이 자유롭게 선택한 태그(tag)를 붙여서 정보를 체계화하는 새로운 분류 체계이다(그림1). 폭소노미 기반의 시스템에서는 사용자(user), 태그(tag), 그리고 리소스(resource) 사이의 관계를 나타내는 3항원소정보(triadic information)를 제공한다. 즉, 대부분의 폭소노미 기반 웹 애플리케이션에서는 사용자가 웹에 존재하는 다양한 리소스(논문, 사진, 동영상, 등)에 태그를 붙여서 3항원소정보를 구성하고, 이를 기반으로 사용자들에게 다양한 서비스를 제공한다. 예를 들면, 사용자들의 북마크를 공유하는 del.icio.us[1]와 북마크뿐만 아니라 논문 등과 웹사이트에 공개된 리소스를 공유하는 BibSonomy[2], 그리고 온라인 사진 관리 및 공유를 위한 Flickr[3]와 동영상 공유를 위한 YouTube[4] 등은 3항원소(사용자, 태그, 리소스)를 토대로 하는 폭소노미 기반 웹 애플리케이션이다.

폭소노미 기반의 웹 애플리케이션 시스템에는 WWW를 이용하는 전 세계의 수많은 사용자들의 다양한 데이터가 축적되어 있으며, 이러한 웹 데이터는 계속적으로 증가·확장·체계화되고, 시시각각으로 변화하는 특성을 가지고



(그림 1) 폭소노미의 3항원소정보 구성의 예

있다. 폭소노미 데이터를 분석하여 유용한 지식·정보들을 추출하기 위한 다양한 연구들[1-4]이 활발하게 진행되고 있다. Ching-man Au Yeung et al.[1]는 애매모호한 태그의 의미를 파악하기 위하여 폭소노미의 구성요소들을 2부 그래프(bipartite graph)를 기반으로 하는 클러스터분석기법을 제안하였다. Robert Jaschke et al.[2]는 tri-concept 데이터 마이닝 기법을 기반으로 폭소노미에 내재되어 있는 암묵적인 사용자 공유정보를 추출하기 위한 기법을 제안하였다. Suk-Hyung Hwang et al.[3]는 폭소노미의 3항원소정보에 대한 계층화된 동치류 분석기법을 제안하였다. 한편, Kim et al.[4]은 폭소노미의 3항원소정보를 형식개념

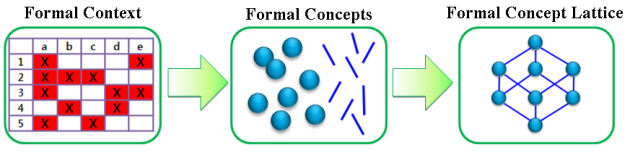
1) http://delicious.com
2) http://www.bibsonomy.org/
3) http://www.flickr.com
4) http://www.youtube.com

분석기법[5,6](FCA : Formal Concept Analysis)을 이용하여 분석하고, 사용자들의 공통태그를 추출하기 위한 기법을 제안하였다.

본 논문에서는, Kim et al.[4]의 후속연구로서, 형식개념분석기법을 기반으로, 사용자, 태그, 리소스들 사이의 3항 관계를 고려한 폭소노미 데이터 마이닝 기법을 제안한다. 구체적으로는, 주어진 리소스에 대한 사용자-태그정보로부터 사용자들의 공통태그정보를 추출하고, 다시 각 공통태그에 대한 사용자-리소스정보로부터 태깅에 사용된 리소스공유정보를 추출함으로써, 특정한 리소스에 대한 공통태그 및 공통사용자 추출뿐만 아니라, 공통태그를 사용하는 사용자그룹과 태깅된 리소스 정보의 클러스터링을 제공한다. 또한, 본 연구에서 제안한 기법을 BibSonomy의 데이터에 적용하여 분석한 실험 결과를 보고한다.

본 논문은 다음과 같이 구성된다. 2장에서는 형식개념분석기법의 기본적인 정의에 대해 설명하고, 3장에서는, 본 연구에서 제안한 형식개념분석기법 기반 폭소노미 데이터 마이닝 기법을 소개한다. 4장에서는, 본 연구에서 제안한 기법을 실제 폭소노미 데이터에 적용하여 실시한 실험결과에 대해 보고하고, 결론 및 향후 연구과제에 대해서 설명한다.

2. 형식개념분석기법(Formal Concept Analysis)



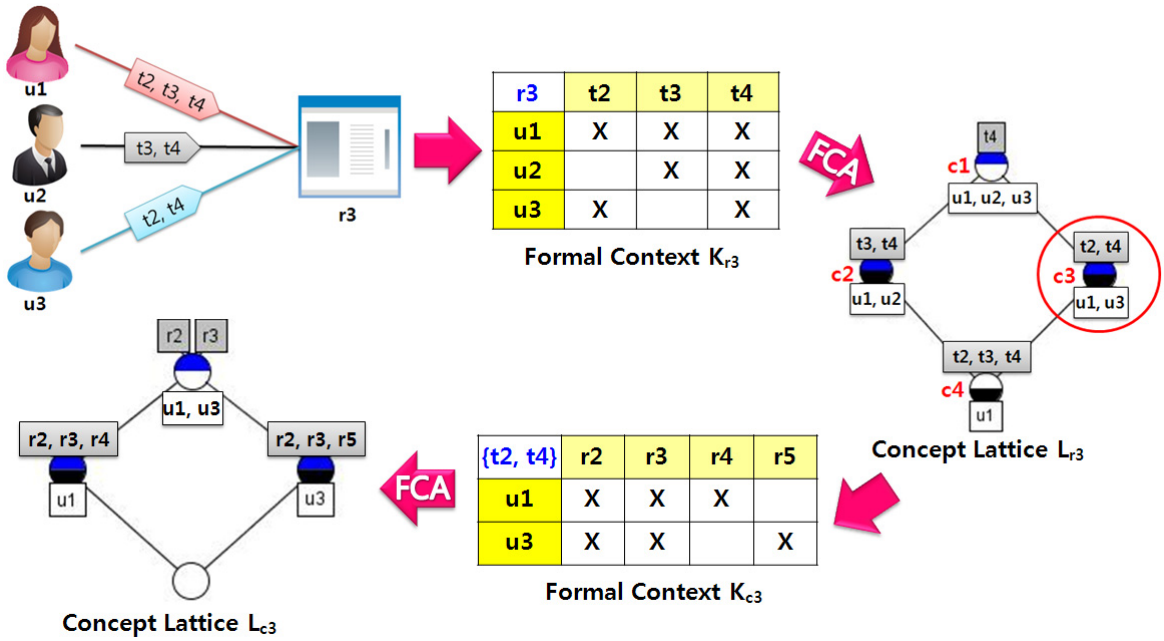
(그림 2) 형식개념분석기법 처리과정

형식개념분석기법은 주어진 데이터로부터 공통속성을 갖는 객체들을 클러스터링 하여 개념(Concept)으로 추출하고 그들 사이의 관계를 토대로 계층화하여 데이터에 내재된 개념들의 구조를 가시화 해주는 데이터분석기법이다(그림 2). 형식개념분석기법을 사용함으로써 실세계의 데이터에 함축된 개념들에 대한 계층구조를 효과적으로 구축할 수 있기 때문에, 현재 의학, 정보과학, 소프트웨어 공학, 등 다양한 분야에서 적용하여 활용되고 있다[6,7].

형식개념분석기법에서는 분석 대상이 되는 데이터가 Formal context라는 일종의 이진데이터 테이블 형태로 주어진다. 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 context K=(G, M, I)에 대하여, O⊆G, A⊆M일 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 같은 객체집합과 속성집합으로 구성된다. 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 개념들의 집합 C와 그들 사이의 상위-하위개념관계 ≤로 이루어진 계층구조L:=(C, ≤)을 개념격자(Concept Lattice 또는 Galois Lattice)라고 부른다.

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. 형식개념분석기법에서는, 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.



(그림 3) 형식개념분석기법 기반 폭소노미 데이터 분석 과정

3. 형식개념분석기법 기반의 폭소노미 데이터 분석

본 장에서는 형식개념분석기법을 기반으로 폭소노미 데이터를 분석하는 방법에 대해서 설명한다.

① 주어진 폭소노미로부터 특정한 리소스 r을 선택하고, 리소스 r을 태깅하고 있는 사용자집합 U와 U의 각 사용자들이 리소스 r을 태깅할 때 사용한 태그들(T)을 추출하여 Formal Context K_r:=(U, T, J)을 생성한다. 단, J = {(u, t, r)| (u, t, r)∈I}.

② 생성된 Formal Context K_r에 형식개념분석기법을 적용하여 리소스 r에 대한 Concept Lattice L_r을 구축한다. Concept Lattice L_r:=(C_r,≤)에는 리소스 r을 어떤 사용자들이 어떤 태그들을 공통적으로 사용하고 있는지에 대한 정보가 개념 단위로 클러스터링 되어 격자 구조로 계층화되어 있다.

③ 제2단계에서 구축된 Concept Lattice L_r에 포함되어있는 개념집합(C_r)의 원소 중에서 관심 있는 공통태그를 포함한 개념 c = (A, B)를 선택한다(A⊆U, B⊆T). 선택된 개념c를 토대로, 아래의 ③-1 또는 ③-2를 수행한다.

③-1 : 공통태그B를 태깅하고 있는 사용자집합A와 A에 포함된 각 사용자들이 태깅한 리소스들의 집합Z를 추출하여 Formal Context K_c:=(A, Z, P)를 생성한다.

③-2 : 사용자집합A가 태깅하고 있는 공통태그B와 A에 의해 공통태그B가 태깅된 리소스들의 집합Z′를 추출하여 Formal Context K_c′:=(B, Z′, P′)를 생성한다.

④ 제3단계에서 구축된 Formal Context K_c(또는 K_c′)를 토대로 형식개념분석기법을 적용하여 개념c에 대한 Concept Lattice L_c(또는 L_c′)를 구축한다. 이 Concept Lattice L_c(또는 L_c′)는, 제1단계에서 선택한 리소스 r에 태깅된 공통태그B(또는 선택한 리소스 r을 태깅하고 있는 사용자집합A)를 기준으로 관련된 사용자들과 리소스들(또는 관련 태그들과 리소스들)에 대한 구조화된 정보를 나타낸다.

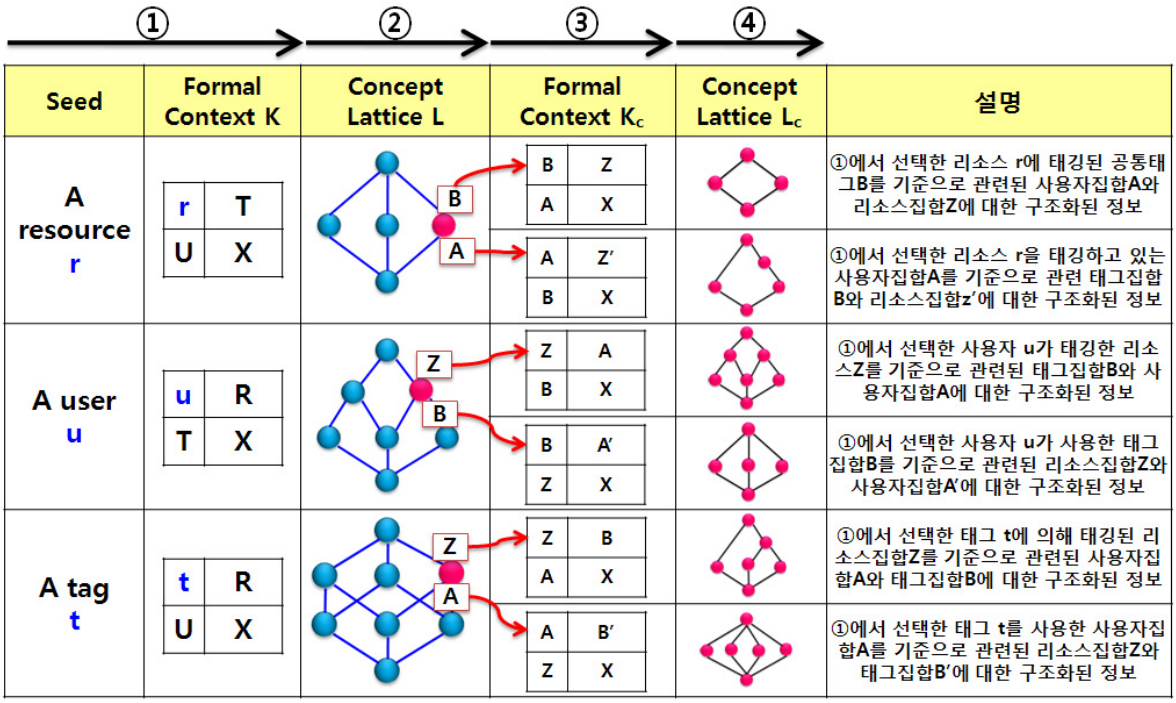
또한, 또다른 분석 방법으로서, 표1에서 정리한 바와 같이 임의의 사용자 u 또는 임의의 태그 t를 선택하여 위와 같은 ① ~ ④의 방법을 적용하여 다양한 측면에서 폭소노미 데이터를 분석할 수 있다.

4. 실험 및 결론

본 장에서는, 실제 폭소노미 데이터에 본 연구에서 제안한 형식개념분석기법 기반의 폭소노미 데이터분석 기법을 적용하여 실시한 실험결과에 대해 보고한다.

Bibsonomy로부터 특정한 리소스 r1[5]을 선택하고, r1을 태깅하고 있는 임의의 사용자 8명과 그들이 리소스 r1을 태깅할 때 사용한 모든 13개의 태그들에 대한 정보를 추출하여 표2와 같이 Formal Context K_{r1}을 생성하였다(①). 표2에 형식개념분석기법을 적용하여 리소스 r1을 태깅하고 있는 사용자들이 공통으로 사용한 태그들에 의해 사용자들을 클러스터링하여 개념들을 추출하고, 그들 사이의 관계를 토대로 그림 4와 같이 Concept Lattice L_{r1}을 구축하였다(②). Concept Lattice L_{r1}에 포함되어 있는 개념들 중에서 "tagging"과 "folksonomy" 태그를 포함한 개념c = ({jboy701, domenico79, rabeeh}, {tagging, folksonomy})를 선택하고, 개념c를 토대로 공통태그"tagging"과 "folksonomy"를 태깅하고 있는 3명의 사용자(jboy701, domenico79, rabeeh)들과 각 사용자들이 공통태그("tagging"과 "folksonomy")를 사용하여 태깅한 31개의 리소스들(r1~r31)에 대한 정보를 추출하여 표3과 같은 Formal Context K_c를 생성하였다(③-1). 표3에 형식개념분석기법을 적용하여 그림5와 같

<표 1> 다양한 측면에서의 폭소노미 데이터 분석

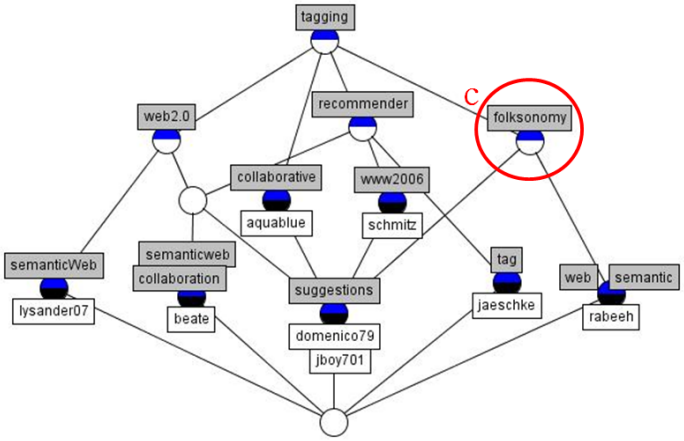


<표 2> Formal Context K_{r1}

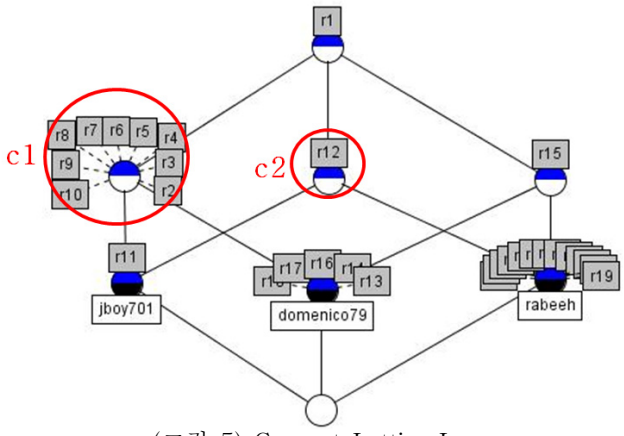
Users \ Tags	www2006	collaboration	collaborative	tagging	tag	semantic	web	suggestions	semanticWeb	web2.0	recommender	folksonomy	semanticweb
beate		X		X						X	X		X
lysander07				X					X	X			
rabeeh				X		X	X					X	
schmitz	X			X							X		
jaeschke				X	X						X		
domenico79	X		X	X				X		X	X	X	
aquablue			X	X									
jboy701	X		X	X				X		X	X	X	

<표 3> Formal Context K_c

users \ resources	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10	r11	r12	r13	r14	r15	r16	r17	r18	r19	r20	r21	r22	r23	r24	r25	r26	r27	r28	r29	r30	r31
jboy701	X	X	X	X	X	X	X	X	X	X	X	X																			
domenico79	X	X	X	X	X	X	X	X	X	X			X	X	X	X	X	X													
rabeeh	X											X			X				X	X	X	X	X	X	X	X	X	X	X	X



(그림 4) Concept Lattice L_{r1}



(그림 5) Concept Lattice L_c

이 개념c에 대한 Concept Lattice L_c를 구축하였다(④). Concept Lattice L_c는, 리소스 r1에 태깅된 공통태그("tagging"과 "folksonomy")를 기준으로 관련된 사용자들과 리소스들에 대한 구조화된 정보를 나타낸다. 예를 들어, Concept Lattice L_c의 개념c1은 사용자 jboy701과 domenico79가 태그"tagging"과 "folksonomy"를 사용하여 공통적으로 태깅한 리소스들(r1, r2, r3, r4, r5, r6, r7, r8, r9, r10)에 대한 클러스터이다. 또한, 개념c2는 사용자 jboy701과 rabeeh가 태그"tagging"과 "folksonomy"를 사용하여 공통적으로 태깅한 리소스가 r1과 r12임을 나타내는 클러스터이다.

본 연구에서 제안한 기법은 특정 태그들을 기준으로 처음 선택하여 입력한 리소스 r1과 관련된 리소스들이 어떤 것들이 있는지, 그리고 그러한 리소스들을 기준이 되는 태그들을 사용하여 어떤 사용자들이 태깅하고 있는지에 대한 정보를 수월하게 추출 할 수 있다. 본 연구에서 제안한 기법을 적용하여 구축된 Concept Lattice Lc를 토대로 특정 리소스와 관련 있는 리소스들, 또는 특정 사용자와 관심사가 같은 사용자그룹, 특정 태그와 관련된 태그들을 추출하고자할 때 유용하게 활용될 수 있다. 현재, 본 논문의 연구 결과를 토대로, 실제 폭소노미 기반의 웹 애플리케이션에 존재하는 다양하고 방대한 양의 태깅 데이터를 분석하기 위한 지원도구를 개발하고 있다.

5)Towards the semantic web : Collaborative tag suggestions (http://www.ibiblio.org/www_tagging/2006/13.pdf)

참고문헌

[1] Ching-man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt, "Understanding the Semantics of Ambiguous Tags in Folksonomies", The International Workshop on Emergent Semantics and Ontology Evolution(ESOE2007) at ISWC/ASWC 2007, pp. 108~121, 2007.

[2] Robert J¨aschke, Andreas Hotho, Christoph Schmitz, Bernhard Ganter, and Gerd Stumme, "Discovering shared conceptualizations in folksonomies", Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 6, Pages 38-53, 2008.

[3] Suk-Hyung Hwang, Yu-Kyung Kang, "Applying Hierarchical Classes Analysis to Triadic context for Folksonomy Mining", 2007 International Conference on Convergence Information Technology (ICCIT'07), pp.103-109, 2007.

[4] Hong-Gee Kim, Suk-Hyung Hwang, Yu-Kyung Kang, Hak-Lae Kim, and Hae-Sool Yang, "An Agent Environment for Contextualizing Folksonomies in a Triadic Context", First KES International Symposium, KES-AMSTA2007, LNAI4496, pp.728-737, 2007. 5.

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

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