http://dx.doi.org/10.3745/KIPSTD.2011.18D.6.423

XOnto-Apriori: 확장된 온톨로지 추론 기반의 연관 규칙 마이닝 알고리즘

이 종 현[†] · 김 장 원[††] · 정 동 원[†††] · 이 석 훈[††] · 백 두 권[††††]

요 약

이 논문에서는 연관 규칙 마이닝 알고리즘의 정확도를 향상시키기 위하여 기존 Onto-Apriori 알고리즘을 확장한 XOnto-Apriori 알고리즘을 제안한다. 기존 알고리즘은 트랜잭션 항목의 식별자만을 비교하여 지지도를 계산하기 때문에 유사한 속성을 가진 항목들간의 관계를 분석하지 못하는 문제점을 지닌다. 이러한 문제점을 해결하기 위해 제안 알고리즘은 온톨로지 추론 기반의 속성 비교를 통해 같은 식별자를 지니지 않는 항목들간의 관계성도 지지도 계산에 반영할 수 있도록 한다. 제안 알고리즘의 규칙 생성 과정을 명확히 서술하기 위해 스마트폰 어플리케이션 추천 시스템을 설계하였으며 이 시스템은 기존 알고리즘 기반의 시스템에 비해 보다 나은 속도와 정확도를 보였다.

키워드 : 연관 규칙 마이닝, 개인화 추천 시스템, 온톨로지 추론, Apriori 알고리즘

XOnto-Apriori: An eXtended Ontology Reasoning-based Association Rule Mining Algorithm

Chonghyeon Lee[†] · Jangwon Kim[††] · Dongwon Jeong[†††] · Sukhoon Lee[††] · Doo-Kwon Baik[††††]

ABSTRACT

In this paper, we introduce XOnto-Apriori algorithm which is an extension of the Onto-Apriori algorithm. The extended algorithm is designed to improve the conventional algorithm's problem of comparing only identifiers of transaction items by reasoning transaction properties of the items which belong in the same category. We show how the mining algorithm works with a smartphone application recommender system based on our extended algorithm to clearly describe the procedures providing personalized recommendations. Further, our simulation results validate our analysis on the algorithm overhead, precision, and recall.

Keywords : Association Rule Mining, Personalized Recommender System, Ontology Reasoning, Apriori Algorithm

1. 서 론

최근 웹 2.0과 시맨틱 웹의 영향으로 온톨로지 태깅을 지원하는 인터넷 서비스들이 증가하고 있다. 온톨로지 태깅은 컨텐트의 분류와 재검색을 용이하게 하며 컨텐트의 의미를 파악하여 상호운용성을 향상시킨다. 인터넷 상의 방대한 정보와 데이터를 통합하여 서비스하기 위하여 온톨로지로 태깅된 정보를 분석하는 다양한 서비스들이 등장하고 있으며, 사용자의 요구와 선호도에 따라 적합한 정보를 제공하는 개인화 추천에 대한 연구가 많이 진행되고 있다[1-3]. 특히 Amazon, yes24와 같은 전자 상거래 분야에서는 이미 개인화 추천이 적용되어 사용자에게 맞춤형 서비스를 제공하고 있으며 그 유용함은 사업적 가치뿐만 아니라 학문적으로도 많은 관심을 불러일으키고 있다[4-5]. 하지만 최근 활성화되고 있는 시맨틱 웹 기술을 활용하여 서로 다른 곳에서 정의한 정보를 분석하여 추천을 제공하는 방법에 대한 연구는 아직 미흡한 상태이다. 그 이유는 온톨로지 태깅으로 분류된 다양한 트랜잭션의 의미 정보들을 반영하여 분석하기 어렵기 때문이다[6].

Onto-Apriori 알고리즘은 온톨로지 태깅된 트랜잭션으로부터 연관규칙을 생성하고 다양한 항목으로부터 구매성향을 능동적으로 분석함으로써 이러한 문제를 해결하였다[7]. 그러나 Onto-Apriori 알고리즘은 온톨로지로 표현된 항목의

※ 이 연구에 참여한 연구자는 '2 단계 BK21 사업'의 지원을 받았으며, 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업입니다(No.2011-0004911). 또한 정보통신산업진흥원의 SW 공학 요소기술 연구개발사업의 결과물임을 밝힙니다.
† 준 회 원 : 고려대학교 컴퓨터 · 전파통신공학과 석사과정
†† 준 회 원 : 고려대학교 컴퓨터 · 전파통신공학과 박사과정
††† 종신회원 : 군산대학교 정보통계학과 교수
†††† 종신회원 : 고려대학교 컴퓨터 · 전파통신공학과 교수
논문접수 : 2011년 8월 23일
수 정 일 : 1차 2011년 10월 14일, 2차 2011년 10월 27일
심사완료 : 2011년 10월 27일

속성이나 클래스간의 관계를 파악하지 않고 지지도를 계산하기 때문에 온톨로지로 명세한 의미 정보들을 충분히 활용하지 못하는 문제점을 지닌다.

이 논문에서는 이러한 문제를 해결하기 위해 Onto-Apriori를 확장한 XOnto-Apriori 알고리즘을 제안한다. XOnto-Apriori 알고리즘은 온톨로지 클래스의 식별자나 이름을 비교하여 규칙을 생성하는 기존 방법과 달리 온톨로지 속성을 비교하여 유사 항목을 추론함으로써 이를 규칙 생성에 반영한다. 이 방법은 서로 다른 자원에서 정의되어 사용되고 있는 온톨로지 리소스의 속성을 비교하여 유사 항목을 하나의 집합으로 재구성함으로써 유사한 속성을 가지고 있음에도 전혀 다른 항목으로 분류되던 문제를 해결할 수 있다. 이는 다양한 곳에서 정의된 온톨로지 트랜잭션에 대한 연관 분석이 가능하며, 더 방대한 양의 자원으로부터 의미를 추출해낼 수 있기 때문에 그 결과로 생성되는 연관 규칙은 개인화 추천의 정확도를 향상시킬 수 있다.

이 논문의 구성은 다음과 같다. 제2장에서 관련연구에 대하여 언급하고 제3장에서 이 논문에서 제안하는 알고리즘을 정의하고 스마트폰 어플리케이션 예제를 통해 알고리즘의 전반적인 절차를 기술한다. 제4장에서는 제안 알고리즘을 평가하기 위하여 XOnto-Apriori 기반의 모델을 구현하여 알고리즘을 검증한다. 마지막으로 제5장에서는 결론 및 향후 연구에 대하여 기술한다.

2. 관련 연구

개인화 추천 시스템은 사용자의 구매 정보나 항목의 정보, 사용자간의 관계 등을 분석하여 선호할 것이라고 예상되는 정보를 제공한다. 이와 같은 시스템을 개발하기 위한 기반 기술로서 데이터 마이닝, 패턴 인식, 정보 필터링 등 다양한 기법들에 대한 연구가 이루어지고 있으며, 그 중 정보 필터링에 대한 연구가 주를 이루고 있다[8]. 개인화 추천을 위한 정보 필터링 기술은 시스템이 사용하는 정보의 특성에 따라 협업적 필터링, 내용 기반 필터링, 규칙 기반 필터링으로 분류할 수 있다[9].

협업적 필터링 기법은 사용자의 개인 정보, 구매 내역 등의 정보를 분석하여 유사한 취향을 가진 이웃의 집합을 찾아 내어 집합 내 이웃들이 선호하는 항목을 추천하는 방법이다[10]. 이 기법은 정보를 분석하는 알고리즘에서 사용자와 항목간 유사도를 모두 계산해야 하기 때문에 방대한 데이터에 적용이 어렵다는 한계를 지닌다.

내용 기반 필터링은 항목 간 유사도를 계산하여 가장 동일한 항목을 추천하는 방식이다[11]. 이 방법은 항목의 속성이 다차원으로 표현될 경우, 협업적 필터링과 마찬가지로 항목의 속성을 추출하는 데 많은 시간이 요구되어 다양한 도메인에 적용하기 어렵다. 또한 단순히 유사도를 계산하여 항목을 비교하기 때문에 유사도 계산 방법에 따라 정확도가 크게 달라지지만, 적용하고자 하는 도메인에 어떤 유사도 비교 알고리즘이 적당한지 결정하기 어렵다는 문제점을 지닌다.

규칙 기반 필터링은 사용자의 성향을 규칙으로 정의하고 이를 상품의 특성과 연결시켜 추천을 제공하는 기법이다[12]. 규칙은 전문가에 의해 미리 정의되기 때문에 다양한 도메인에 적용하기 어려우며, 여러 곳에서 정의된 방대한 통합 트랜잭션에 적용할 수 없다.

이러한 문제를 해결하기 위한 방법으로 온톨로지 추론 기반의 연관 규칙 마이닝 알고리즘이 제안 되었으며 대표적으로 xPMiner, Onto-Apriori 알고리즘이 있다[13]. xPMiner는 온톨로지 기반의 웹 사용자 마이닝 알고리즘으로서 Apriori 기반의 연관 규칙 마이닝을 통해 추천을 제공한다. xPMiner는 온톨로지를 적용한 도메인 지식 구축 활용은 특정 도메인에 종속하지 않은 분석이 가능하지만, 온톨로지의 의미 정보를 활용하지 않아 추천 정확도가 다소 떨어지는 단점을 가지고 있다. Onto-Apriori 알고리즘은 온톨로지 추론을 기반으로 항목 비교를 수행함으로써 불필요한 항목들을 후보 집합에 포함하지 않는다. 하지만 지지도를 계산할 때 항목을 온톨로지 클래스의 식별자만을 비교하기 때문에, 실제 같은 속성을 갖는 항목이더라도 같은 식별자를 가지고 있지 않다면 계산에 포함시키지 않는다. 이러한 지지도 계산은 온톨로지로 표현된 트랜잭션의 속성과 항목간의 관계 등을 반영하지 못하는 문제점을 지닌다. 이 논문은 온톨로지 속성을 비교하여 같은 항목임을 추론함으로써 보다 개선된 결과를 얻을 수 있는 XOnto-Apriori 알고리즘을 제안하여 이러한 문제를 해결하고자 한다.

3. XOnto-Apriori 알고리즘

XOnto-Apriori 알고리즘은 Onto-Apriori를 확장한 알고리즘으로서, 기존 방법과 달리 지지도(Support) 및 신뢰도(Confidence)를 계산할 때 온톨로지 추론을 활용하여 같은 속성의 항목은 서로 다른 식별자로 정의되어 있더라도 결과에 반영하는 방법이다. 기존 Onto-Apriori 알고리즘은 전체 트랜잭션 중 항목 X와 Y의 온톨로지 식별자를 비교하여 지지도와 신뢰도를 계산한다. 그러나 이와 같은 지지도 계산 방법은 온톨로지로 표현된 속성들과 데이터들간의 관계를 충분히 반영할 수 없다. 또한 X와 Y가 같은 속성을 갖는 사실상 같은 항목이더라도 온톨로지 식별자만으로 다른 항목으로 분류하여 지지도 계산에서 후보 항목으로 선정되지 않는 문제점을 지닌다. 이러한 문제점은 웹 상의 여러 리소스로부터 트랜잭션을 수집하여 계산할 때 극대화되며, 이를 해결하기 위해 XOnto-Apriori 알고리즘에서는 아래 수식과 같이 지지도와 신뢰도를 X와 Y의 속성으로 비교하여 계산함으로써 알고리즘의 정확도를 향상시킨다. 각 계산 과정의 추론은 ReasoningCount(i, T) 함수에서 이루어지며 빈발 항목 집합들 중 최소 지지도(*minsup*) 이상의 지지도를 만족하는 항목들을 빈발 항목 집합에 포함시키고 최종 빈발 항목 집합으로부터 최소 신뢰도(*mincorf*)를 갖는 항목들을 추천 엔진이 사용할 규칙으로 생성한다.

$$\text{Support} = \frac{\{\text{Reasoning}(X, T) \cup \text{Reasoning}(Y, T)\}}{n}$$

$$\text{Confidence} = \frac{\{\text{Reasoning}(X, T) \cup \text{Reasoning}(Y, T)\}}{\text{Reasoning}(X, T)}$$

이 절에서는 제안 알고리즘을 활용한 추천 과정의 장점을 서술하기 위하여 스마트폰 어플리케이션 추천을 예로 사용한다. 트랜잭션 데이터는 온톨로지로 표현된 엔티티를 사용하며, (그림 1)은 스마트폰 어플리케이션 온톨로지 스키마의 예를 나타낸다. 연관 관계를 분석하기 위해서 클래스의 속성을 명세한 온톨로지 스키마가 필요하며 같은 속성을 비교하여 지지도 및 신뢰도를 계산한다.

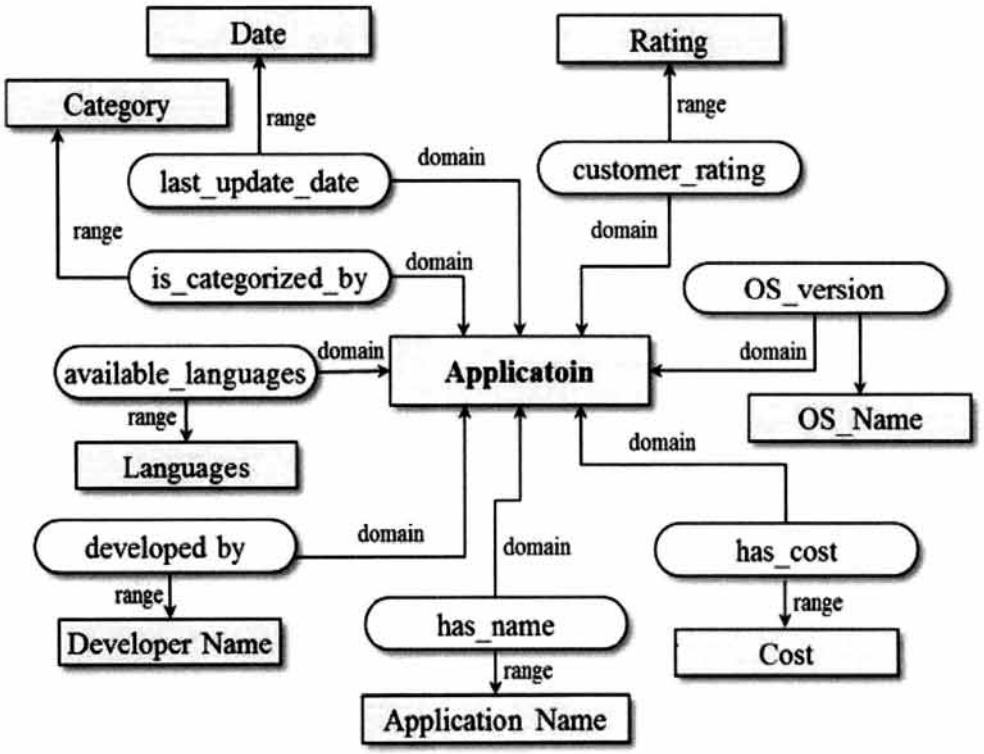
스키마를 기반으로 작성된 트랜잭션 항목의 예는 (그림 2)와 같다. 각 어플리케이션 인스턴스는 위의 스키마를 참조하여 작성되었으며, 트위터, 페이스북 등의 어플리케이션을 정의된 스키마 속성을 사용하여 버전, 실행 환경, 사용 가능 언어 등의 정보를 표현한다. 이러한 항목들은 <표 1>과 같이 트랜잭션 목록으로 나타낼 수 있으며 사용자가 다운로드 받은 어플리케이션의 집합을 포함한다. t_n은 알고리즘을 적용한 예를 설명하기 위한 사용자들의 트랜잭션 집합의 일부이다. 클래스의 이름을 식별자로 <application>과 같이 온톨로지로 표현된 어플리케이션을 나타내며 이는 사용자가 다운로드한 어플리케이션을 나타낸다.

트랜잭션 목록을 분석하기 위한 제안한 알고리즘은 <표 2>, <표 3>, <표 4>와 같다. XOnto-Apriori(T) 함수는 트랜잭션 집합 T의 항목들로부터 부분집합을 생성하며 빈발 항목 집합 F를 생성하는 함수이다. 제안 알고리즘에서는 기존 방법과 달리 온톨로지 속성을 비교하기 위한 ReasoningCount(i, T) 함수를 추가하였으며 이는 부분집합을 생성하여 지지도를 계산할 때 항목들간의 관계를 추론하여 반영한다.

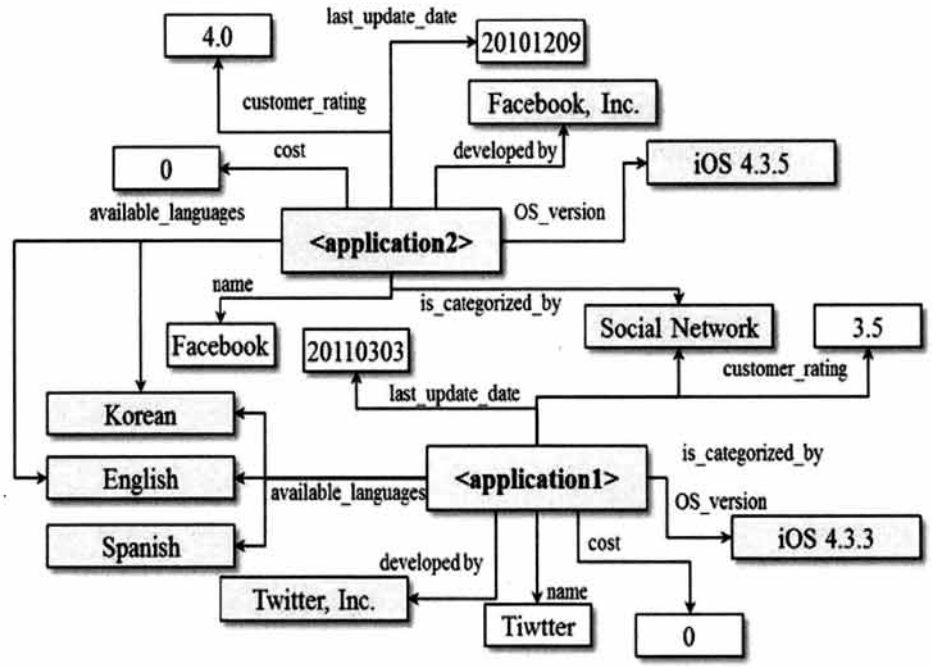
XOnto-Apriori(T) 함수는 먼저 트랜잭션 집합 T를 스캔하여 트랜잭션 리스트를 초기 후보 항목 집합인 C_1에 할당한다. 1단계 빈발 항목 집합인 F_1을 얻기 위하여 생성된 C_1을 ReasoningCount(i, T) 함수에 대입하여 각 어플리케이션의 속성값을 비교함으로써 지지도를 계산한다.

〈표 1〉 예제 트랜잭션 집합

트랜잭션	스마트폰 어플리케이션
t_1	{<트위터>, <푸딩카메라>, <카카오톡>, <올댓페이스북>}
t_2	{<트위터>, <페이스북>, <어썸노트>, <Informant>}
t_3	{<어썸노트>, <Informant>, <범프>}
t_4	{<트위터>, <범프>}
t_5	{<트위터>, <페이스북>, <어썸노트>, <Informant>}
t_6	{<푸딩카메라>, <마이피플>, <카카오톡>}
t_7	{<트위터>, <푸딩카메라>, <페이스북>, <어썸노트>}
t_8	{<어썸노트>, <Informant>, <올댓페이스북>}



(그림 1) 스마트폰 어플리케이션 온톨로지 스키마의 예



〈그림 2〉 스마트폰 어플리케이션 온톨로지 엔티티의 예

〈표 2〉 XOnto-Apriori 알고리즘

XOnto-Apriori (*T*)

1:$C_1 \leftarrow$ init-pass(T);
2:$F_1 \leftarrow \{f \mid f \in C_1$, ReasoningCount($f$, T)/$n \geq minsup\}$;
3: for ($k = 2$; $F_{k-1} \neq \emptyset k$++) do
4:$C_k \leftarrow$ candidate-gen(F_{k-1});
5:for each transaction $t \in T$ do
6:for each candidate $c \in C_k$do
7:if reasoning(c, T) is contained in t then
8:c.ReasoningCount++;
9:endfor
10:endfor
11:$F_k \leftarrow \{c \in C_k \mid c$.ReasoningCount($f$, T)/$n \geq minsup\}$
12:endfor
13: return $F \leftarrow \cup_k F_k$

〈표 3〉 온톨로지 추론을 이용한 항목 비교

ReasoningCount(*i*, *T*)

1: $I_c \leftarrow$getCategory(i, *domain ontology*)
2: for each transaction $t \in T$ do
3: $t_c \leftarrow$getCategory(t, *domain ontology*)
4: if I_c is t_c then
5: return *true*
6: endfor
7: return *false*

초기화 한 C_1으로부터 XOnto-Apriori-Candidate-Gen(F_{k-1}) 함수를 통해 지지도를 계산하여 생성한 F_1은 <표 5>와 같다. 각 단계에서 생성한 후보항목 c_n은 트랜잭션 모음의 항목들 중 원소의 개수가 하나인 모든 부분집합이다. 기존 알고리즘과 달리 후보 항목의 집합 C_k의 부분집합을 점검할 때 ReasoningCount(i, T) 함수를 사용하여 식별자가 다르더라도 같은 카테고리로 분류되는 것을 추론하여 삭제되지 않도록 처리한다. 예제에서 <페이스북>과 <올댓페이스북>은 OS만 다를 뿐 페이스북 기반의 유사한 속성을 가진 항목이므로 c_{10}과 같이 논리 기호인 +로 묶어 새로운 집합으로 생성하여올댓페이스북>을 다른 항목으로 분류하여 계산하므로 각 지지도가 최소 지지도보다 작아 3 단계 후보 항목으

〈표 4〉 후보 항목 집합 생성

XOnto-Apriori-Candidate-Gen(F_{k-1})

1:$C_k \leftarrow \emptyset$
2:for all $f_1, f_2 \in F_{k-1}$
3: with $f_1 = \{i_1, \ldots, i_{k-2}, i_{k-1}\}$
4: and $f_2 = \{i_1, \ldots, i_{k-2}, i'_{k-1}\}$
5: and $i_{k-1} < i'_{k-1}$do
6: $c \leftarrow \{i_1, \ldots i_{k-1}, i'_{k-1}\}$;
7: $C_k \leftarrow C_k \cup \{c\}$;
8: for each (k−1)-subset s of c do
9: if (ReasoningCount(s, F_{k-1}) $\notin F_{k-1}$) then
10: delete c from C_k
11: endfor
12: endfor
13: return C_k

〈표 5〉 2단계 후보 항목 지지도

후보 항목	지지도
c_1 = <트위터>	5/8
c_2 = <푸딩카메라>	3/8
c_3 = <페이스북>	3/8
c_4 = <어썸노트>	5/8
c_5 = <Informant>	4/8
c_6 = <범프>	2/8
c_7 = <마이피플>	1/8
c_8 = <카카오톡>	2/8
c_9 = <올댓페이스북>	2/8
c_{10} = {<페이스북>, <올댓페이스북>} → <페이스북 + 올댓페이스북>	5/8

로 선택하지 않지만, 하나의 집합으로 생성함으로써 지지도를 계산할 때 같은 항목으로 추론하여 5/8의 지지도를 갖는다. 이러한 과정은 <페이스북>과 <올댓페이스북>을 3단계 후보 항목으로 생성될 수 있도록 지지도에 포함시켜 Apriori 알고리즘의 원리인 하강 속성을 유지하게 한다. 이 과정에서 후보 항목을 생성하기 위해서 i_{k-1}과 i'_{k-1}만 다른 것을 조합한다. 이는 다른 항목이 두 가지 이상인 것을 조합하여 후보 항목 생성할 경우 k-1 부분집합에 속하지 않아 후보로 선택되지 않기 때문이다.

생성된 2단계 후보 항목인 C_2의 지지도를 최소 지지도인 0.5와 비교하여 {<트위터>}, {<페이스북>}, {<범프>}, {<트위터>, <페이스북>}와 같은 2차 빈발 항목집합 F_2를 추출하고 이를 기반으로 후보 생성 알고리즘을 한번 더 수행하면 <표 6>과 같은 3단계 후보 항목 지지도 C_3을 얻을 수 있다. <페이스북 + 올댓페이스북>을 C_3로 추출할 수 있었기 때문에 c_1, c_4, c_5와 같은 후보 항목을 얻을 수 있다.

〈표 6〉 3단계 후보 항목 지지도

후보 항목	지지도
c_1 = <<트위터>, <페이스북 + 올댓페이스북>>	4/8
c_2 = <<트위터>, <어썸노트>>	3/8
c_3 = <<트위터>, <Informant>>	2/8
c_4 = <<페이스북 + 올댓페이스북>, <어썸노트>>	3/8
c_5 = <<페이스북 + 올댓페이스북>, <Informant>>	3/8
c_6 = <<어썸노트>, <Informant>>	4/8

C_3로부터 최소 지지도 이상의 지지도를 갖는 빈발 항목집합 F_3로부터 생성하는 부분집합은 더 이상 빈발 항목으로 추출되는 집합이 없으므로 F_3은 최종 빈발 항목집합이 되며 이를 기반으로 생성한 연관 규칙 집합 R은 <표 7>과 같다. R의 모든 연관 규칙은 최소 신뢰도인 0.5보다 높은 신뢰도를 갖기 때문에 추천을 위한 최종 연관 규칙으로 사용한다.

〈표 7〉 생성된 연관 규칙

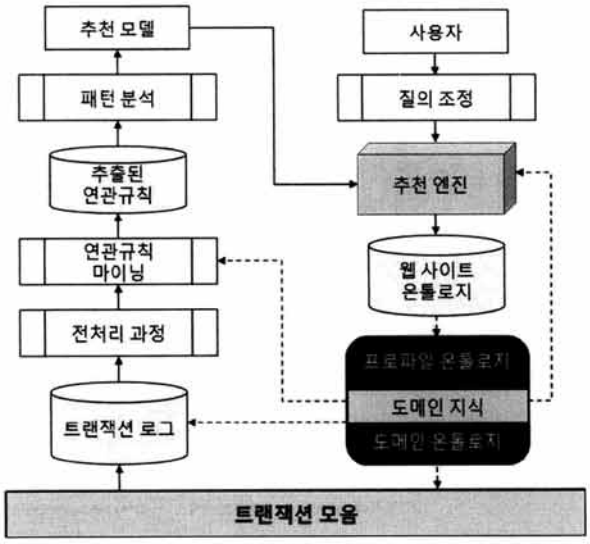
연관 규칙	신뢰도
r_1 = <트위터> → <페이스북 + 올댓페이스북>	4/5
r_2 = <페이스북 + 올댓페이스북> → <트위터>	4/5
r_3 = <어썸노트> → <Informant>	4/5
r_4 = <Informant> → <어썸노트>	4/4

4. 평 가

이 장에서는 제안한 XOnto-Apriori 알고리즘을 Onto-Apriori 알고리즘, 최적화된 Apriori 알고리즘, 온톨로지 기반의 트랜잭션 배열을 적용한 xPMiner와 비교 평가를 통해 제안 알고리즘이 정확도와 성능 측면에서 기존 방법보다 효과적임을 보인다. 비교 평가를 위한 실험은 Windows 7 Enterprise K, Intel® Core™ 2 Duo E6750 2.66GHz CPU, 3.25 GB Memory, Java™ SE Development Kit 6, Jena Semantic Web Framework 2.6.3으로 구성된 환경에서 수행하였다.

4.1 XOnto-Apriori 알고리즘 기반 시스템 아키텍처

제안 알고리즘을 기반으로 개인화 추천을 제공하기 위해 이 논문에서는 (그림 3)과 같이 시스템을 설계하여 구현하였다. 추천 시스템의 마이닝 프로세스는 전처리, 연관 규칙 마이닝, 패턴 분석, 추천의 과정으로 이루어진다. 사용자에게 적합한 추천을 제공하기 위해 서버의 트랜잭션 모음으로부터 로그를 가져와 전처리 과정을 수행한다. 전처리 과정은 웹 페이지를 세션 별로 분류하는 세션화, 데이터 필드의 불필요한 참조를 제거하는 데이터 클리닝, 웹 서버의 로그 파일을 융합하는 데이터 퓨전 작업을 통해 트랜잭션의 속성들을 마이닝 할 수 있도록 식별하는 작업이다. 제안 알고리즘

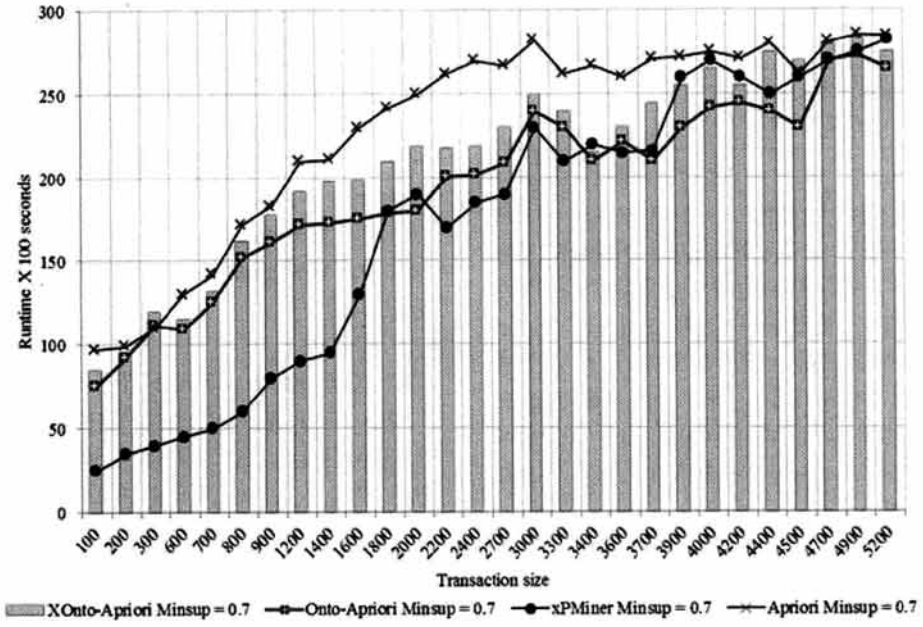


(그림 3) XOnto-Apriori 기반의 개인화 추천 시스템 아키텍처

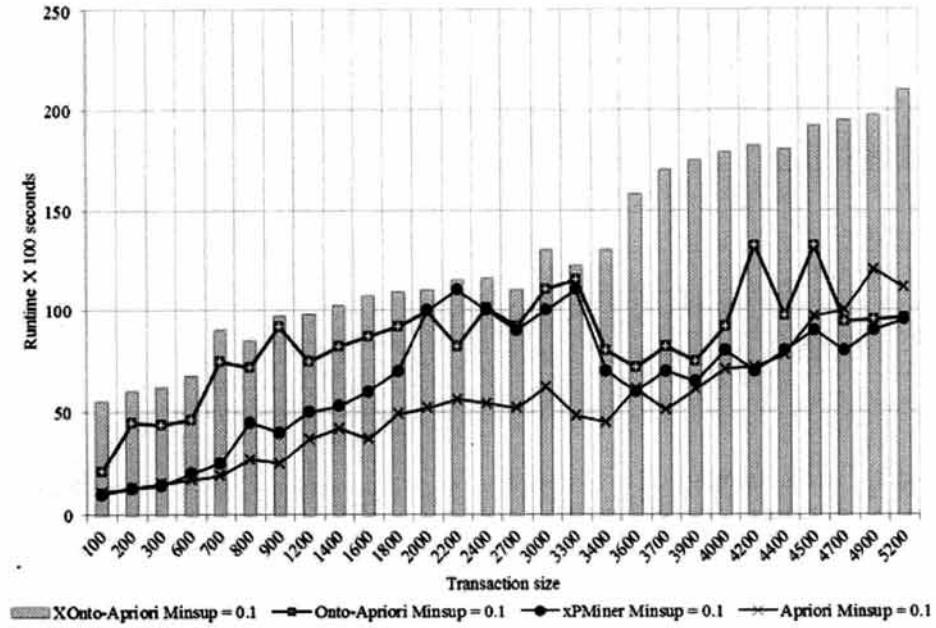
은 전처리 과정을 거친 트랜잭션을 활용하여 연관 규칙을 생성한다. 추출한 연관 규칙의 패턴을 분석하여 규칙으로 이루어진 추천 모델을 생성하면 추천 엔진은 사용자에게 맞는 추천을 추론해 낼 수 있다. 사용자는 SPARQL 기반의 질의를 사용하여 이러한 모델로부터 추천을 제공받을 수 있다. 연관 규칙 마이닝에 사용되는 알고리즘뿐만 아니라 추천 엔진, 트랜잭션 데이터는 온톨로지로 표현되어 관리되기 때문에 도메인 지식을 통해 각기 다른 곳에서 가져온 데이터를 의미 기반으로 처리할 수 있다.

4.2 알고리즘 처리 속도 비교

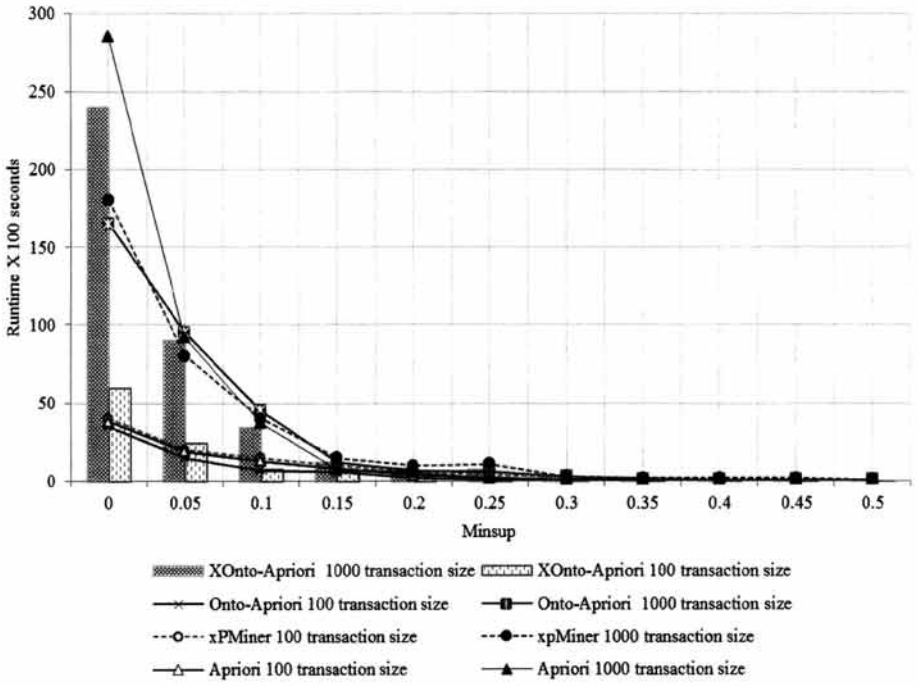
제안 알고리즘의 적용 범위를 비교 분석하기 위하여 처리 속도에 영향을 미치는 최소 지지도와 트랜잭션의 항목 집합 개수를 증가시키며 처리 속도를 측정하였다. 비교 알고리즘으로 Onto-Apriori 알고리즘, 최적화된 Apriori 알고리즘, 온톨로지 기반의 트랜잭션 배열을 적용한 xPMiner를 구현하여 비교 평가를 수행하였다. 알고리즘에서 사용된 항목 집합은 온톨로지로 표현된 스마트폰 어플리케이션을 사용하였으며 항목 집합의 개수의 변화에 따른 처리 속도 측정 결과



(그림 4) 트랜잭션 크기에 따른 처리 속도(*minsup*=0.7)



(그림 5) 트랜잭션 크기에 따른 처리 속도(*minsup*=0.1)



(그림 6) 최소 지지도에 따른 처리 속도

는 최소 지지도 값에 따라 (그림 4), (그림 5)와 같다. 최소 지지도를 낮게 설정할수록 생성되는 후보 항목 집합이 많아지기 때문에 처리 속도가 저하되며, 항목 집합의 개수를 증가시키면 후보 항목 집합을 생성하는 과정에서 비교 횟수가 증가하기 때문에 오버헤드가 증가함을 확인할 수 있다. 특히 Onto-Apriori 알고리즘에 비해 대체적으로 낮은 속도를 보였으나 추론 횟수는 부분집합 생성에 비해 작으므로 시간 복잡도가 동일하여 큰 차이를 보이지 않는다. 이는 XOnto-Apriori 알고리즘의 온톨로지 속성 비교 과정에서 유사 항목을 추론하는 작업이 추가 되었기 때문이다. 그러나 0.7과 같이 높은 최소 지지도를 설정하였을 경우 추론 과정 또한 감소하여 Onto-Apriori 알고리즘, xPMiner와 유사한 성능을 보였으며 Apriori 알고리즘이 Apriori-gen 과정의 부하 때문에 가장 낮은 성능을 보였다.

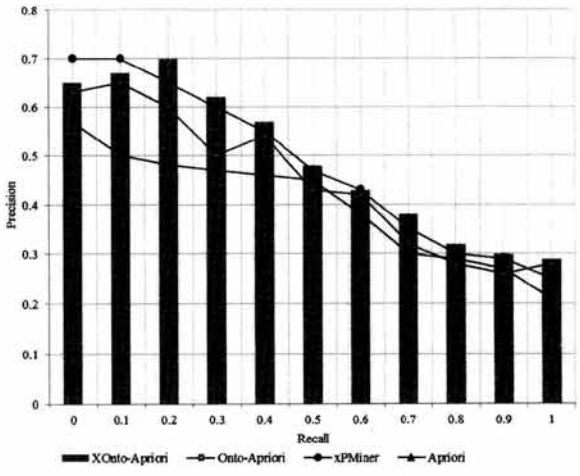
최소 지지도의 변화에 따른 각 알고리즘의 처리 속도를 측정한 결과는 (그림 6)과 같다. 최소 지지도가 낮을수록 후보 집합의 항목 개수가 많아지므로 연관 규칙 추출 속도가 현저히 떨어지게 된다. 앞서 수행한 평가와 동일하게, 최소 지지도가 낮을 때 XOnto-Apriori 알고리즘이 다른 알고리즘에 비해 낮은 수행시간이 소요됨을 확인할 수 있다. 하지만 최소 지지도를 높일수록 처리 속도 차이는 미비해지며, XOnto-Apriori 알고리즘은 최소 지지도를 높게 설정하더라도 온톨로지 추론을 통해 의미 있는 연관 규칙을 추출해 낼 수 있으므로 처리 속도에서의 문제점을 해결할 수 있다. 또한 오히려 속도 측면에서도 좋은 결과를 보이기도 하는데, 이는 추론을 통해 한가지 항목으로 묶어 항목을 생성하는 것이 부분집합으로 생성되는 것보다 빠르기 때문이다.

4.3 정확도 및 재현율 분석

이 논문에서 제안한 알고리즘의 규칙 생성 결과를 아래 수식과 같은 정확도(Precision) 및 재현율(Recall)로 평가한다. 정확도는 생성된 규칙 중에서 적절한 규칙의 수를, 재현율은 전체 적절한 규칙 중에서 추출된 규칙의 수를 나타낸다. 정확도와 재현율은 반비례 관계이기 때문에 정확도를 높이기 위해 최소 지지도를 높이면 적절한 규칙의 대부분을 추출하지 못한다[14].

$$\text{Precision} = \frac{|\{\text{relevant rules}\} \cap \{\text{retrieved rules}\}|}{|\{\text{retrieved rules}\}|}$$

$$\text{Recall} = \frac{|\{\text{relevant rules}\} \cap \{\text{retrieved rules}\}|}{|\{\text{relevant rules}\}|}$$



(그림 7) 모델 별 정확도 및 재현율

평가를 위한 세 가지 알고리즘을 추천 모델로 구현하여 정확도에 따른 재현율을 평가한 그래프는 (그림 7)과 같다. 정확도를 향상시키기 위해 최소 지지도를 높일 경우 C_n의 항목 개수가 감소하기 때문에 재현율이 떨어지게 된다. Apriori 모델의 경우, 규칙 생성 속도에만 목적을 둔 알고리즘이기 때문에 가장 낮은 정확도와 재현율을 보였다. 반면 XOnto-Apriori 모델은 온톨로지 속성 추론으로 유사 항목까지 규칙 생성에 반영시키기 때문에 가장 좋은 정확도와 재현율을 보임을 확인할 수 있다.

4.4 F-Measure 분석

정확도와 재현율은 반비례 관계이기 때문에 이 두 값을 모두 고려하여 알고리즘을 평가하기 위해 두 값의 조화평균인 F-Measure를 함께 측정하였다[15]. F-Measure의 계산식은 아래와 같으며, β는 1로 설정하여 정확도와 재현율의 비율을 동등하게 하였다.

$$F_\beta = (1+\beta^2)\cdot\frac{precision\cdot recall}{\beta^2\cdot precision + recall}$$

제안 모델을 비롯한 세 가지 모델의 F_1-Measure 조화평균을 계산한 결과는 <표 8>과 같다. 평과 결과 T_3, T_6, T_9, T_{10}을 제외한 모든 트랜잭션에서 가장 높은 정확도를 나타내었으며 온톨로지 속성을 통해 유사 항목을 많이 추출한 T_8에 대해 가장 큰 성능 차이를 보였다. T_3과 같이 식별자가 같지만 속성이 다른 특별한 경우에는 기존 알고리즘보다 낮은 정확도를 보였으며, T_9와 T_{11}는 온톨로지 식별자가 잘 정의되어 Onto-Apriori 알고리즘과 제안 방법이 같은 결과를 나타내었다. 하지만 제안 알고리즘이 기존 방법들과 비교하여 가장 높은 F_1-Measure 평균을 얻음으로써 제안 방법이 중요한 규칙만을 선별하고 결과의 정확도를 높일 수 있음을 확인하였다.

〈표 8〉 F1-Measure 값 비교

	F_1-Measure			
Transaction Set	XOnto-Apriori	Onto-Apriori	Apriori	xPMiner
T_1	0.7257	0.6524	0.5786	0.5588
T_2	0.3600	0.3310	0.3252	0.3059
T_3	0.1017	0.1121	0.1267	0.0863
T_4	0.2324	0.2324	0.3859	0.2433
T_5	0.2176	0.1625	0.1987	0.1794
T_6	0.1082	0.0725	0.1267	0.0950
T_7	0.1940	0.1250	0.1813	0.1745
T_8	0.4128	0.3750	0.1218	0.1631
T_9	0.5063	0.5063	0.4192	0.5395
T_{10}	0.1940	0.1750	0.2801	0.1560
T_{11}	0.2176	0.2176	0.2057	0.2028
T_{12}	0.1794	0.1699	0.1388	0.0902
T_{13}	0.5063	0.4502	0.4831	0.3721
Average	0.3043	0.2125	0.2747	0.2436

5. 결 론

이 논문에서는 이러한 문제를 개선하기 위해 온톨로지 엔지니어링 기술을 연관규칙 생성 알고리즘에 접목한 XOnto-Apriori 알고리즘을 제안하였다. 온톨로지 추론을 통한 속성 비교는 서로 같은 식별자를 갖지 않는 항목이더라도 규칙에 영향을 줄 만큼 유사하면 지지도 계산에 포함함으로써 규칙으로 생성될 수 있도록 하여 개인화 추천의 정확도를 향상시킨다. 알고리즘을 평가하기 위해 제안 알고리즘 기반의 추천 모델을 설계하여 시스템으로 구현하였으며 Apriori 기반의 다른 알고리즘들과 비교하여 정확도 및 재현율 면에서 더 나은 성능을 보임을 검증하였다.

연관규칙 생성 알고리즘은 후보 항목 집합을 생성하는 과정에서 가능한 모든 부분집합을 생성하기 때문에 높은 시간 복잡도를 지닌다. 그러나 개인화 추천을 위해 연관 관계 분석은 반드시 필요한 과정이므로 추천 시스템은 추천의 정확도를 향상시키기 위하여 연관규칙 생성 알고리즘을 사용한다. 이 논문은 이러한 알고리즘의 정확도와 재현율을 향상시키기 위한 연구이며, 제안 알고리즘뿐만 아니라 기존 알고리즘도 같은 시간 복잡도를 갖기 때문에 트랜잭션의 항목 개수가 증가함에 따라 처리 시간이 증가함을 확인할 수 있다. 이러한 연관규칙 생성 알고리즘의 시간 복잡도 문제를 개선하기 위한 방법은 향후 연구로 수행할 예정이다.

참 고 문 헌

[1] A. Felferning, M. Mandl, J. Tihonen, M. Schubert, and G. Leitner, "Personalized user interfaces for product configuration", ACM Proceeding of the 14th international conference on Intelligent user interfaces, pp.317-320, 2010.

[2] Y. Wu, Y. Chen, and A.L.P. Chen, "Enabling personalized recommendation on the Web based on user interests and behaviors", Research Issues in Data Engineering Proceedings. Eleventh International Workshop on, pp.17-24, 2001.

[3] S.Y. Ho and S.H. Kwok, "The Attraction of Personalized Service for Users in Mobile Commerce: An Empirical Study", ACM SIGecom Exchanges, Vol.3, No.4, pp.10-18, 2003.

[4] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: item-to-item collaborative filtering", Internet Computing, IEEE Computer Society, pp.76-80, 2003.

[5] J. Choi, H. Lee, and Y. Kim, "The Influence of Social Presence on Evaluating Personalized Recommender Systems", Pacific Asia Conference on Information Systems(PACIS), AIS Electronic Library (AISeL), 2009.

[6] V. Tamma, "Semantic Web Support for Intelligent Search and Retrieval of Business Knowledge", IEEE Intelligent Systems, Vol.25, No.1, pp.84-88, 2010.

[7] C. Lee, S. Lee, J. Kim, D. Baik, "A Personalized Clothing Recommender System Based on the Algorithm for Mining

Association Rules", Journal of Korea Society for Simulation, Vol.19, No.4, pp.59-66, 2010

[8] X. Y. Su and Taghi M. Khoshgoftaar, "A survey of collaborative ˉlerting techniques. Advances in Artiˉcial Intelligence", Vol.2009, No.4, 2009.

[9] M. J. Pazzani, "A Framework for Collaborative, Content-Based and Demographic Filtering", Artificial Intelligence Review, Vol.13, No.5-6, pp.393-408, 1999.

[10] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews", In Proc. of ACM CSCW'94 Conference on Computer Supported Cooperative Work, pp.175-186, 1994.

[11] J. Herlocker, J. Konstan, A. Borchers, and J. Riedl, "An Algorithm Framework for Performing Collaborative Filtering", In Proc. of ACM SIGIR'99, 1999.

[12] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering", In Proc. of the 14th Conference on Uncertainty in Artificial Intelligence, 1998.

[13] R. Missaoui, P. Valtchev, C. Djeraba, and M. Adda, "Toward Recommendation Based on Ontology-Powered Web-Usage Mining", IEEE Internet Computing, Vol.11, No.4, pp.45-52, 2007.

[14] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves", In ICML S06: Proceedings of the 23rd international conference on Machine learning, pp.233 - 240, 2006.

[15] T. Y. Chen, F. Kuo, and R. Merkel, "On the Statistical Properties of the F-Measure", Quality Software Fourth International Conference on (QSIC'04), pp.146-153, 2004.

이 종 현

e-mail : momoline@korea.ac.kr
2009년 고려대학교 컴퓨터정보학과(학사)
2010년~현 재 고려대학교 컴퓨터·전파통신공학과 석사과정
관심분야 : 데이터 마이닝, 시맨틱 웹, 온톨로지 엔지니어링, 개인화 추천 시스템

김 장 원

e-mail : ikaros1223@korea.ac.kr
2005년 상명대학교 소프트웨어공학과(학사)
2005년 한국과학기술연구원(KIST) 위촉연구원
2008년 고려대학교 컴퓨터학과(석사)
2008년~현 재 고려대학교 컴퓨터·전파통신공학과 박사과정
관심분야 : 온톨로지, 시맨틱 웹, GIS, 데이터베이스, 메타데이터 등

정 동 원

e-mail : djeong@kunsan.ac.kr
1997년 군산대학교 컴퓨터과학과(이학사)
1999년 충북대학교 전산학과(이학석사)
2004년 고려대학교 컴퓨터학과(이학박사)
1998년 전자통신연구원 위촉연구원
1999년~2000년 ICU 부설 한국정보통신교육원 GIS 분원 전임강사
2000년~2001년 (주)지구넷 부설 연구소 선임연구원
2002년~2005년 라임디미디어 테크놀로지 부설 연구소 연구원
2004년~2005년 고려대학교 정보통신기술연구소 연구조교수
2005년 Pennsylvania State University PostDoc
2002년~2004년 TTA 표준화위원회-데이터연구회(SG08.02) 특별위원
2004년~현 재 TTA 표준화위원회-메타데이타 표준화 프로젝트그룹(PG406) 위원
2005년~현 재 군산대학교 정보통계학과 교수
2006년~현 재 데이터관리서비스 전문위원회(ISO/IEC JTC 1/SC 32 Mirror Committee) 위원
2008년~현 재 지리정보 전문위원회(ISO/TC 211 Mirror Committee) 위원
2009년~현 재 TTA 표준화위원회-NGIS 프로젝트그룹(PG409) 위원
2010년~현 재 인터넷윤리실천협회 이사
2010년~현 재 ICDL Korea 교수위원
2010년~현 재 전북지역 과학기술정보협의회 위원
2010년~현 재 한국과학기술정보연구원 자문위원
2010년~현 재 한국컴퓨터교육학회 이사
관심분야 : 데이타베이스, 시맨틱 웹, 시맨틱 GIS, 유비쿼터스 컴퓨팅, 시맨틱 모바일 서비스, 클라우드 컴퓨팅

이 석 훈

e-mail : brandon@korea.ac.kr
2009년 고려대학교 전자 및 정보공학부(학사)
2009년~2011년 고려대학교 컴퓨터·전파통신공학과 (공학석사)
2011년~현 재 고려대학교 컴퓨터·전파통신공학과 박사과정
관심분야 : 시맨틱 웹, 온톨로지, 데이터마이닝, 메타데이타 레지스트리

백 두 권

e-mail : baikdk@korea.ac.kr

1974년 고려대학교 수학과(학사)

1977년 고려대학교 산업공학과(석사)

1983년 Wayne State Univ. 전산학과(석사)

1985년 Wayne State Univ. 전산학과(박사)

1989년~2007년 (사)한국정보과학회 (이사/평의원/부회장)

1986년~현 재 고려대학교 컴퓨터·전파통신공학과 교수

1991년~현 재 (사)한국시뮬레이션학회 (이사/부회장/감사/회장/고문)

1991년~현 재 ISO/IEC JTC1/SC32 전문위원회(위원장)

2001년~현 재 (사)도산아카데미(원장)

2002년~2004년 고려대학교정보통신대학(초대학장)

2004년~2005년 (사)정보처리학회(부회장)

2009년~2010년 고려대학교 정보통신대학 학장

관심분야 : 메타데이터, 소프트웨어공학, 데이터공학, 컴포넌트 기반 시스템, 메타데이터 레지스트리, 프로젝트 매니지먼트 등