모바일 디바이스 상에서의 특이성 탐지를 위한 베이지안 추론 모델

(A Bayesian Inference Model for Landmarks Detection on Mobile Devices)

황 금 성 † 조 성 배 †† 이 종 호 †††
(Keum-Sung Hwang) (Sung-Bae Cho) (Jong-Ho Lea)

요 약 모바일 디바이스에서 얻을 수 있는 로그 데이타는 의미 있고 실속 있는 다양한 개인 정보를 담고 있다. 그러나 메모리 용량과 연산 능력의 제한, 분석의 어려움으로 인해 이러한 정보들은 무시되고 있는 것이 일반적이다. 모바일 환경의 이러한 어려움을 극복하기 위해 로그 데이타를 분산된 모듈에서 분석하여 사용자에게 의미 있는 정보인 특이성을 탐지하는 새로운 방법을 제안한다. 제안하는 방법은 불확실한 상황에서의 추론 정확도를 향상시키기 위해 베이지안 확률 접근 방법을 채택하고 있다. 새로운 협력적 모듈형 기술은 모바일 디바이스의 제한된 자원을 가지고 효율적으로 연산하기 위해 베이지안 네트워크를 모듈로 나눈다. 인공 데이타와 실제 데이타를 이용한 실험에서 인공 데이타의 경우 약 84%의 정확률과 약 76%의 재현률을 보였으며, 실제 데이타에서는 부분 일치를 포함하여 약 89%의 일치율을 보였다.

키워드 : 모바일 디바이스 로그, 특이성 추출 모델, 모듈화된 베이지안 네트워크

Abstract The log data collected from mobile devices contains diverse meaningful and practical personal information. However, this information is usually ignored because of its limitation of memory capacity, computation power and analysis. We propose a novel method that detects landmarks of meaningful information for users by analyzing the log data in distributed modules to overcome the problems of mobile environment. The proposed method adopts Bayesian probabilistic approach to enhance the inference accuracy under the uncertain environments. The new cooperative modularization technique divides Bayesian network into modules to compute efficiently with limited resources. Experiments with artificial data and real data indicate that the result with artificial data is amount to about 84% precision rate and about 76% recall rate, and that including partial matching with real data is about 89% hitting rate.

Key words : mobile device log, landmark extraction model, modular Bayesian network

1. 서 론

모바일 환경은 여러 가지 면에서 기존의 컴퓨팅 환경과 다른 특성을 가진다. 먼저, 모바일 디바이스는 통화 기록, SMS, 사진, MP3, GPS 등과 같은 다양한 정보를 다루고 수집할 수 있다. 또한, 모바일 디바이스는 개인성이 강한 장비이므로 개인의 기호나 성향에 따라 적응되어 특화될 수 있다. 그리고, 모바일 디바이스는 사용자가 항상 휴대하기 때문에 사용자의 일상정보를 효과적으로 수집하고 분석하여 사용자에게 도움을 줄 수 있다[1]. 이러한 모바일 디바이스의 특성은 사용자 편의를 위한 다양한 서비스 제공의 가능성을 열어 주었고, 최근에는 Nokia[1] 등의 기업과 많은 연구자들에게 연구 및 개발의 대상으로 관심을 받고 있다. 특히, 최근 인간 중심의 기술로서 활발하게 연구되고 있는 컨텍스트 어웨어 기술은 모바일 환경에서 더욱 많은 활용 가능성을 가지고 있다[2,3]. 따라서, 지능형 통화 서비스[4], 메시징 서비스[5], 모바일 로그의 분석과 수집·관리[1,6-11]와 같은 다양한 지능형 서비스가 연구되고 있다.

하지만, 모바일 디바이스는 PC에 비해서 적은 메모리

† 학생회원 : 연세대학교 컴퓨터과학과
yellowg@sclab.yonsei.ac.kr
†† 종신회원 : 연세대학교 컴퓨터과학과 교수
sbcho@sclab.yonsei.ac.kr
††† 정 회 원 : 삼성종합기술원 CIL 전문연구원
john.lea@samsung.com
논문접수 : 2006년 9월 29일
심사완료 : 2006년 12월 4일

1) Nokia LifeBlog 프로젝트, http://www.nokia.com/lifeblog

용량, 적은 CPU 처리용량, 작은 화면 크기, 불편한 입력 인터페이스, 제한된 배터리 용량 등의 한계를 가지고 있으며, 변화가 심한 실세계 환경에서 작동되기 때문에 더욱 능동적이고 효과적인 적응 기능이 요구되는 점 등이 개발의 어려운 요인이다[12].

본 논문에서는 모바일 환경에서 수집된 로그 정보를 효과적으로 분석하고 효율적으로 고수준의 컨텍스트 및 특이성(Landmark, 특별히 기억에 남을 정보)을 추출하기 위한 방법을 제안한다. 제안하는 방법은 모바일 환경에서 발생하는 다양한 불확실성(① 실생활의 불규칙성, ② 사용자 의도 및 감정의 불확실성, ③ 센서의 불확실성, ④ 인과관계의 불확실성)을 효율적으로 다루기 위해 베이지안(Bayesian) 확률 모델[13]을 채택하였으며, 베이지안 확률 모델이 모바일 환경에서 효과적으로 동작할 수 있도록 하기 위해 협력적 모듈형 베이지안 네트워크(Cooperative modular Bayesian network) 모델을 제안한다.

최근 로그 정보를 분석하여 향상된 서비스를 제공하고자 하는 시도가 활발하다. A. Krause 등은 모바일 디바이스에서 수집된 센서 및 로그 정보를 클러스터링하고 사용자의 기호를 반영하는 컨텍스트에 부합하도록 학습시켜 사용자의 상황 예측 및 서비스 제공을 하였다[6]. 이때 컨텍스트에 대한 서비스 선택 방법으로 베이지안 네트워크(BN)를 사용하였다. 하지만 이 연구에서는 전통적인 베이지안 네트워크 모델을 이용하여 좁은 도메인의 분류 문제에만 적용하고 있다. 도메인이 큰 경우, 전통적인 베이지안 네트워크 모델은 높은 복잡도의 연산을 요구하기 때문에 모바일 디바이스에서는 어렵다.

E. Horvitz 등은 베이지안 네트워크 기술을 기반으로 PC의 로그 데이타에서 학습된 인간의 인식 활동 모델을 만들고 이를 통해 랜드마크를 발견하고 추론하는 방법을 제안하였다. 하지만 이 방법은 모델의 크기가 매우 크고 복잡하여 용량 및 성능이 제한된 모바일 환경에서 그대로 사용하기엔 제약이 많다. 따라서 모바일 환경에 좀더 최적화된 방법이 요구된다. 2003년에 모바일 환경에서 사용자에 대한 방해 상황(interruptability)을 나타내는 컨텍스트를 추론하는 모델을 개발하였으나, 비교적 단순한 상황에 대해서 2개의 BN을 사용한 연구였고, 큰 규모의 컨텍스트 추론에 대해서는 다루지 않았다[14].

2. 모바일 로그 데이타 분석 및 컨텍스트 생성

본 장에서는 일상생활 속에서 모바일 디바이스에 저장되는 사용자 정보를 수집하는 방법을 정리하고 컨텍스트를 획득하는 방법을 제안한다. 컨텍스트는 정보의 빈도, 지속시간, 발생 간격을 통계적으로 분석하고 임팩트 수치를 계산하여 구한다. 얻어진 컨텍스트는 사용자가 하루 동안 경험한 일들을 추론하기 위한 증거로 이용된다.

2.1 사용자 정보 수집

표 1은 모바일 디바이스에서 수집되는 사용자 정보의 내용을 설명한다. GPS 정보로부터 사용자가 방문한 장소에 대한 정보를 얻을 수 있으며, Call과 SMS 정보로부터 사용자가 통화한 내역과 빈도를 구할 수 있다. 인터넷으로부터 얻을 수 있는 날씨 정보는 사용자가 하루 중에 경험한 기분이나 상태에 영향을 미칠 수 있다. MP3 음악은 사용자의 감성에 영향을 주고 사진을 찍은 내역은 사용자가 기억하고 싶은 일이 있음을 알 수 있다. 사진을 본 내역은 사용자가 즐겨 보는 사진이나 그림에 대한 정보를 제공해 준다.

표 2는 사용자 정보의 수집 방법과 주기를 설명한다. 사용자 정보는 수집되는 정보의 종류에 따라서 수집 주기가 다르다. 사진보기나 MP3 듣기 내역의 경우에는 이미지 뷰어나 MP3 플레이어 프로그램을 이용할 때마다 수집된다. SMS와 Call, 사진 찍은 내역과 날씨 정보는 하루에 한번 수집된다. GPS 위치 정보와 충전상태 정보는 매 1초 마다 수집된다.

2.2 컨텍스트 생성

수집된 사용자 정보가 추론 모델이나 서비스에 이용되기 위해서는 컨텍스트를 생성할 필요가 있다. 특히 GPS 정보는 모바일 디바이스에 저장된 기록만으로는

표 1 수집되는 로그 정보

로그 종류	얻을 수 있는 정보
GPS	위도, 경도, 이동속도, 진행방향, 날짜, 시간
Call	상대방 전화번호, 송신/수신/부재 여부, 통화 시작/종료시간
SMS	상대방 전화번호, 송신/수신 여부, 발신/수신 시간
사진보기	사진파일명, 사진보기 시작한 시간, 사진 닫은 시간
사진	사진파일명, 사진 파일 생성날짜
날씨	날씨, 시정(km), 전운량(%), 현재기온(°C), 불쾌지수(%), 체감온도(°C), 강수량(mm), 적설(cm), 습도(%), 풍향, 풍속(m/s), 해면기압(hPa)
MP3	노래제목, 시작시간, 종료시간, MP3가 시작한 위치
충전상태	현재 충전량, 충전 중인지 여부, 현재 시각

표 2 로그 수집 방법과 수집 주기

로그 종류	수집 방법	수집 주기
GPS	GPS 모듈로부터 수집	매 1초마다 로그를 남김
Call	저장된 통화 내역을 수집	매 1일 마다 로그를 남김
SMS	저장된 SMS 내역을 수집	매 1일 마다 로그를 남김
사진보기	이미지 뷰어를 사용하여 수집	매 수행 시 로그를 남김
사진	사진파일 생성내역 수집	매 1일 마다 로그를 남김
날씨	인터넷에서 날씨 정보를 수집	매 1일 마다 로그를 남김
MP3	MP3 플레이어로부터 수집	매 수행 시 로그를 남김
충전상태	백그라운드 프로그램으로 수집	매 1초마다 로그를 남김

사용자의 위치를 판단할 수 없고 기록되는 정보량이 많아 추론 모델이나 서비스에서 직접 이용하기가 어렵다. 따라서 통계 정보와 임팩트를 이용하여 컨텍스트를 생성하는 방법을 제안한다. 또한 GPS 정보로부터 사용자의 위치를 분석하고 필요한 정보를 추출하는 방법을 제안한다.

(1) 통계적 분석과 임팩트 분석

로그 정보에서 빈번하게 발생하거나 오래 지속되는 사건, 혹은 평소에 거의 발생하지 않는 사건이라면 사용자에게 의미 있는 일이 될 수 있다. 따라서 '하루 동안 발생한 빈도(횟수)', '하루 동안 지속된 시간의 총량', '마지막 발생 후 지난 시간'에 대해 통계적 분석을 한다.

사용자에게 의미 있는 일은 단순히 빈도나 지속 시간만으로 판단하기는 어렵고, 사건의 집중도가 사용자에게 더욱 의미 있을 가능성이 높다. 임팩트는 특정 사건이 발생한 빈도에 대한 집중도를 의미하는 것이다. 표 3에서는 임팩트가 증가하는 기준과 감소하는 기준을 보여주고 있다. 일정 시간 내로 이벤트가 연속적으로 발생되면 임팩트 수치는 점점 증가하게 된다. 즉, 임팩트 수치로 이벤트의 빈도와 집중도가 높은 시점을 파악할 수 있다.

(2) 위치 분석

그림 1은 본 논문에서 사용된 GPS 위치 정보 분석 과정이다. GPS 위치 정보를 도, 분, 초에서 X, Y 좌표로 환산하고 레이블링 된 장소인 경우에는 방문한 장소의 목록을 생성한다. 이때, 연속된 데이타인 경우에는 시작 시간, 끝 시간을 합쳐서 통합하여 방문한 전체 시간이 계산된다.

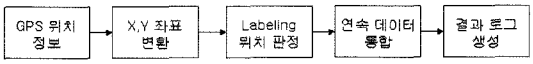

그림 1 GPS 정보를 이용한 사용자 위치 분석 과정

사용자가 방문한 장소들을 레이블링하여 위치 컨텍스트를 생성할 때 장소 인식을 위해 다음과 같은 2가지 방식을 사용하였다. 장소가 넓은 지역일 경우에는 다각형 방식을, 좁은 지역일 경우에는 중심점 방식을 사용하였다. ①다각형 방식은 지역을 다각형으로 표시하고 사용자가 그 다각형의 내부에 있을 경우 그 장소를 방문한 것으로 판단한다. 넓은 지역의 장소 판단에 사용되었다. ②중심점 방식은 건물이나 지역의 중심으로부터의 거리가 오차허용범위 이하일 경우 건물이나 지역을 방문한 것으로 판단한다. 좁은 지역의 장소 판단에 사용하였다. 중심점 방식의 지역 인식을 위해 지역을 다각형으로 표시하고 다각형을 이루는 점들(x_1,y_1)~(x_n,y_n)의 좌표에서 다각형의 중심점(x_m, y_m)의 좌표를 수식 (1)과 같이 계산한다. 중심점으로부터의 거리가 수식 (2)의 건물의 반경 R보다 작은 경우에는 방문한 것으로 판단한다. 이때, R은 중심점에서 다각형의 한 점까지의 거리와 R_{err} (건물에 들어가기 전에 GPS가 끊기는 경우를 고려한 오차허용범위)를 합하여 계산한다. 본 논문에서 사용한 R_{err} 값은 0.3초(GPS 거리 단위)이다.

$$x_m = \sum_{i=1}^{n} \frac{x_i}{n}, \qquad y_m = \sum_{i=1}^{n} \frac{y_i}{n} \tag{1}$$

$$R = \sqrt{(x_m - x_1)^2 + (y_m + y_1)^2} + R_{err} \tag{2}$$

다각형 방식으로 모든 지역의 장소 인식이 가능하지

표 3 임팩트 수치의 증감. 여기에서 임팩트 수치 변화 시간은 로그 데이타에 따라 다르게 선택되었다.

종류	임팩트 수치 증가	임팩트 수치 감소
GPS	If (GPS Event) Impact++	If (Impact>0 AND 매1시간) Impact--
Call	If (Call Event) Impact++	If (Impact>0 AND 매1시간) Impact--
SMS	If (SMS Event) Impact++	If (Impact>0 AND 매20분) Impact--
사진보기	If (사진 보기) Impact++	If (Impact>0 AND 매5분) Impact--
사진	If (사진 찍기) Impact++	If (Impact>0 AND 매30분) Impact--
MP3	If (음악 듣기) Impact++	If (Impact>0 AND 매30분) Impact--

만, 오차 범위에 대한 고려가 쉽지 않기 때문에 면적이 좁은 장소의 경우에는 인식이 잘 되지 않는다. 넓은 지역에 대해 중심점 방식을 사용한 경우에는 거짓 긍정(False Positive) 오류가 크게 발생한다. 또한 좁은 지역에서 다각형 방식을 사용한 경우 GPS 좌표의 오차를 극복하지 못하여 인식률이 떨어진다. 그림 2는 실제 사용자가 방문한 좁은 장소에 대한 인식률을 비교한 결과이다. 중심점 방식이 대체적으로 좋은 성능을 보였으며, 오차범위가 0.3초일 때 가장 좋은 결과를 보였다.

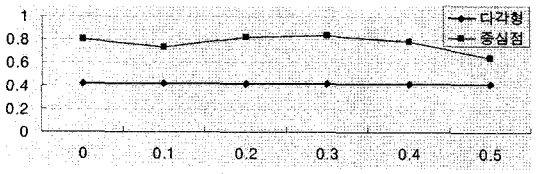

그림 2 좁은 지역에서의 오차 허용 범위(x축, 단위 초)에 따른 위치 판단의 정확도(y축) 비교

실험에 사용되는 지역의 종류를 일반화하기 위해 다음과 같은 대분류, 소분류, 장소 특성 항목으로 나누어서 사용하였다. 대분류 장소는 소분류 장소를 포함할 수 있는 넓은 범위를 의미한다. 소분류 장소는 대분류 장소에 포함되며 장소 특성을 가지는 항목이다. 이때, 대분류 장소(16개)는 {학교, 자연, 운동, 공연, 집, 전자상가, 유흥, 교통, 쇼핑, 외식, 관람, 휴식, 종교, 행사, 일터, 병원}이고, 소분류 장소(53개)는 {고등학교, 고등학교 운동장, 대학교 강의동, 도서관, 학교 식당, 동아리방, 교문, 노천 극장, 집, 사무실, 전철역, 공항, 기차역, 선착장, 병원, 백화점, 길거리 농구장, 헬스 클럽, 테니스장, 골프장, 운동장, 체육관, 아이스 스케이트장안, 스키장, 농구 경기장, 축구 경기장, 야구 경기장, 경마장, 실내 수영장, 실외 수영장, 절, 교회, 성당, 콘서트장, 연극/뮤지컬, 극장, 결혼식장, 묘지, 숲속, 공원, 동물원, 수족관, 박물관, 식물원, 유원지/놀이공원, 찜질방, 산 정상, 목장, 항구, 해수욕장, 커피점, 패스트푸드점, 식당}이며, 장소 특성은 {실내, 실외}이다.

3. 모바일 로그에서의 특이성 탐지

모바일 환경에서 수집된 로그를 분석하여 특이성을 추출하는 과정은 그림 3과 같다. 모바일 디바이스에서 수집된 다양한 로그는 전처리를 거쳐 특이성 추론 모듈에 의해 특이성이 결정된다. 이때, 패턴 인식이나 간단한 논리 규칙에 의한 특이성 추론은 1차 전처리 모듈에서 수행하고, 복잡한 확률적 추론은 BN에 의한 특이성 추론 모듈에서 수행하였다. 이는 규칙만으로도 추출이 가능한 특이성의 경우 규칙 모듈에서 처리함으로써 BN

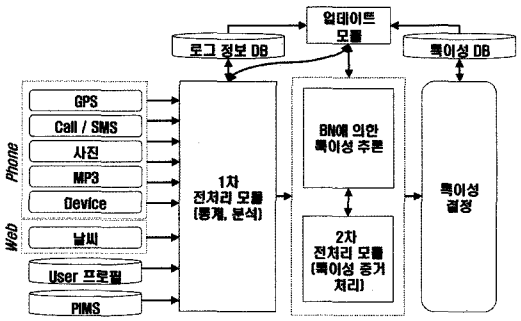

그림 3 모바일 로그에서의 특이성 추출 과정

의 복잡도를 줄이고 BN에서 증거 정보로 활용할 수 있도록 하기 위함이다. BN에 의해 추론된 특이성을 분석하고 특이성 증거로 활용하기 위해 2차 전처리 모듈이 사용되었다. 이는 제안하는 협력적 모듈형 베이지안 네트워크 구조가 1차, 2차 추론 과정을 거칠 때 1차 추론의 결과를 2차 추론의 증거로 사용하기 때문에 이를 고려한 구조이다.

베이지안 네트워크는 노드의 연결 관계를 표현하는 방향성 비순환 그래프(DAG: directed acyclic graph) 형태이며, 이 구조에 따라 정의된 조건부 확률 테이블(CPT: conditional probability table)에 의해 적은 비용으로 많은 확률 관계를 효율적으로 표현 및 계산할 수 있는 모델이다[13-15]. 그림 4는 실제로 설계된 BN의 예를 보여주며, DAG 구조와 노드 이름, 상태 이름, 추론된 확률값을 보여준다.

3.1 협력적 모듈형 베이지안 네트워크

본 논문에서 제안하는 베이지안 네트워크는 기존의 방법과 다른 점이 크게 두 가지이다. 첫 번째, 확률 추론 모델을 분할된 도메인에 따라 모듈화하여 사용한다(그림 5). 베이지안 네트워크의 특성상 노드와 연결의 수가 많아질수록 더 많은 컴퓨팅 성능을 요구하게 된다. 특히, 하나의 노드에 여러 원인 노드가 연결될 경우 복잡도가 $O(k^N)$ (k는 상태의 수, N은 부모의 수)에 비례하기 때문에 BN이 작을수록 모바일 환경에 유리하다.

두 번째, 모듈화된 BN에서의 상호 인과성을 반영하기 위해 그림 6과 같은 2단계의 추론 과정을 거친다. 이때 특이성 증거를 좀더 정확히 반영하기 위해 가상 증거 기술을 사용하였다. 이 방법은 확률적인 증거를 반영하기 위해 가상 노드를 추가하여 노드의 확률값(CPV: conditional probability value)을 통해 증거의 확률을 적용하는 방법이다[17].

가상 증거 기술은 그림 7처럼 주어진 증거가 확률적인 특성을 가진 경우 이를 반영하기 위해 가상 노드를 자식노드로 정의하여 가변적인 확률 테이블을 사용하는

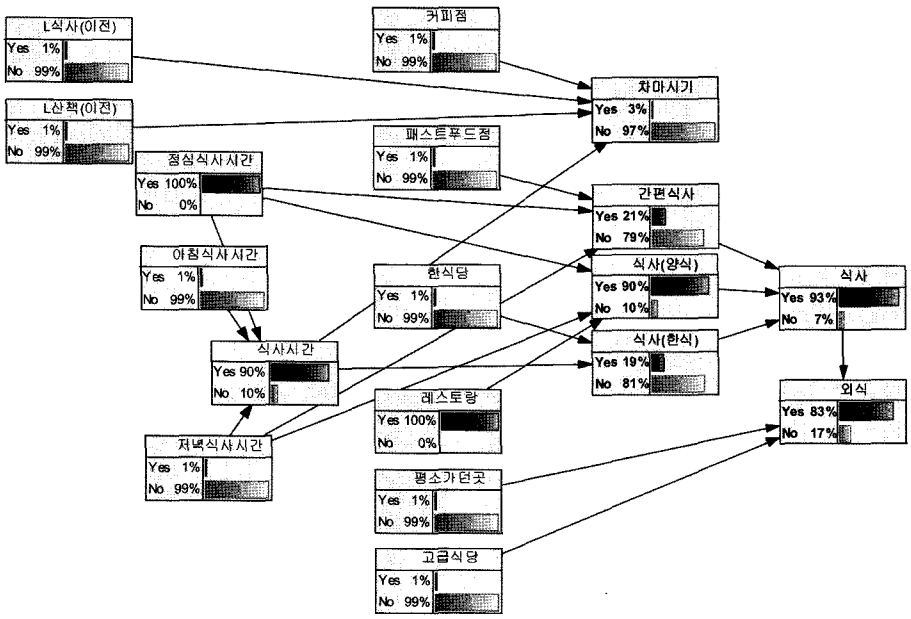

그림 4 외식지역 관련 특이성 추론을 위해 설계된 '외식지역 행동' 특이성 추론 BN

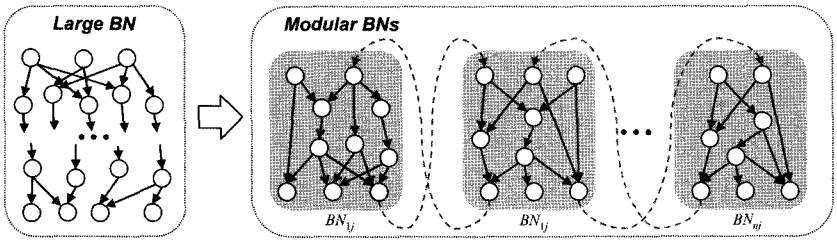

그림 5 모듈화된 베이지안 네트워크. 그림에서 점선은 가상 연결에 의해 연결된 노드를 표현한다. 가상 연결은 노드 간의 확률적인 인과관계를 반영하기 위해 CPT 확률값을 조정하여 증거를 반영한다.

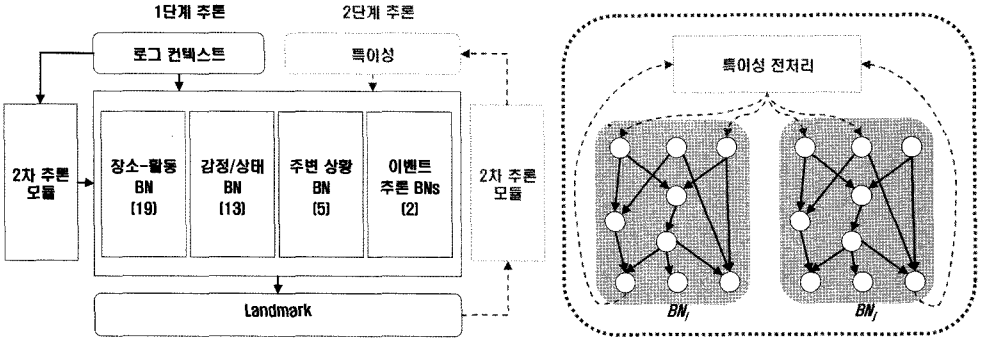

그림 6 협력적 모듈형 베이지안 네트워크 추론 과정. 괄호 안의 숫자는 포함된 BN의 수를 의미한다. 점선은 2단계 추론 과정을 나타내며, 오른쪽 그림은 1단계 추론 결과가 2단계에서 여러 BN의 증거로 사용됨을 보이고 있다.

방법이다. 본 논문에서는 베이지안 네트워크 구조의 원형을 유지하기 위해 가장 오른쪽 형태의 가상 노드를 제안하여 사용하였다. 이 방법은 초기 확률값을 포기하는 대신 확률 추가되는 노드 없이 가상 증거를 반영할 수 있는 방법이며, 루트 노드에서만 사용이 가능한 방법이다.

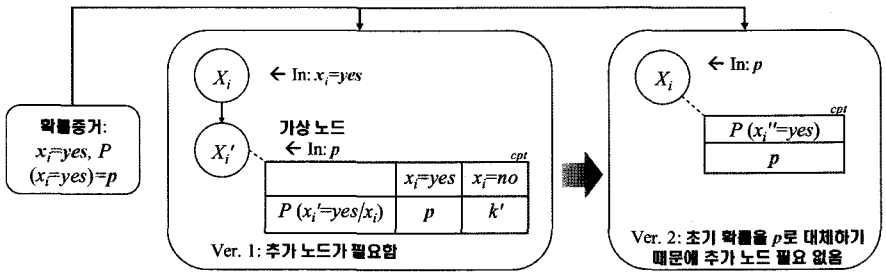

그림 7 가상 증거 기술. 주어진 증거가 확률적인 값을 가지고 있을 경우 사용한다. Ver 2를 제안하여 사용하였다.

예를 들어 BN_1의 구조가 {A→B→C}이고 BN_2의 구조가 BN_1의 모듈화된 구조인 {A→B, B→C}인 경우, 증거 A가 주어졌다면, 제안하는 가상 증거 기술과 체인 룰을 적용하여 노드 C에 대한 믿음(belief)을 다음과 같이 계산할 수 있다[13].

$$BN_1\text{'s } Bel(C) = P(A,C) = P(C|B)P(B|A)P(A) \quad (3)$$

$$BN_2\text{'s } Bel(B) = P(A,B) = P(B|A)P(A) \quad (4)$$

$$BN_2\text{'s } Bel(C) = P(B,C) = P(C|B)\ P(B) = P(C|B)$$

$$P(A,B) = P(C|B)\ P(B|A)P(A) \quad (5)$$

$$\therefore BN_2\text{'s } Bel(C) = BN_1\text{'s } Bel(C) \quad (6)$$

이때, 가상 증거 기술의 가정에 의해 $P(B) = Bel(B)$이다. 수식(3~6)을 통해 모듈화된 가상 증거기술을 사용한 2개의 BN이 통합된 1개의 BN과 동일한 추론 결과를 냄을 알 수 있다.

3.2 특이성 부가 정보 분석

BN 추론 모듈에서 추출되는 특이성의 확률값을 통해 특이성의 신뢰도를 알 수 있으며, 연결관계를 통해 특이성이 추론된 배경 및 인과 관계를 알 수 있다. 이때 인과성의 강도를 구분 짓고 계산하기 위해 NoisyOR 가중치를 계산하여 사용하였다. NoisyOR 가중치는 설계 및 학습 비용을 줄이기 위한 베이지안 확률 테이블 계산 방법의 하나인 NoisyOR BN 모델에서 사용되는 원인별 조건부 확률의 연결 강도 S_i를 의미하며 수식 (7)과 같이 계산된다.

$$S_i = (p_i / 0.5) - 1.0 \quad (7)$$

여기서 p_i는 원인 x_i가 활성화된 경우의 조건부 확률 값을 의미한다.

$$p_i = \Pr(y \mid \bar{x}_1, \bar{x}_2, \cdots, x_i, \cdots, \bar{x}_{n-1}, \bar{x}_n) \quad (8)$$

3.3 복잡도 비교

실험에 사용된 BN은 총 39개로서, 장소 별 행동 추론 BN(19개,{관람, 교통, 모임, 바쁨, 병원, 사진, 쇼핑, 연락, 외식, 운동, 유흥, 음악, 이동행동, 일터, 자연, 종교, 집, 학교, 휴식}), 감정·상태 추론 BN(13개,{놀람, 당황, 더움, 배고픔, 심심, 아픔, 우울, 짜증, 추움, 취함, 즐거움, 피곤, 화남}), 주변 상황 추론 BN(5개, {공간, 그룹상태, 날씨, 시간, 디바이스상태}), 이벤트 추론 BN(2개,{기념일, 행사})이다. 이 BN들은 표 4에서 보이는 바와 같이 638개의 노드와 623개의 링크, 4,205개의 CPV로 구성되어 있다.

하나의 BN으로 모델을 구성한 경우에는 표에서와 같이 462개의 노드를 가진다. 모듈화된 BN보다 노드수가 적은 이유는 중복된 노드가 없기 때문이다. 하지만 부모 노드의 수와 CPV의 크기는 증가하기 때문에 복잡도가 크다. 즉, 모듈화된 BN은 평균 16.6개의 노드와 107.8개의 CPV를 이용한 추론 연산이 39번 수행되는데 비해 단일화된 모델은 469개의 노드와 4,869개의 CPV를 가진 추론 연산이 1번 수행되므로 효율성이 떨어진다. BN 연산 복잡도가 노드의 수와 CPV의 수에 비례한다고 가정하면, 복잡도 $O'(NN \times NCPV)$는 단일 BN 모델의 경우 469×4,869=2,283,561인데 비해 모듈화된 BN의 경우는 39×16.6×107.8=69,790이므로 단일화된 경우에 대해 약 3%의 복잡도만 가지고 약 33배의 효율성을 가진다. 실제 모바일 환경에서는 메모리 용량과 계산 성능의 제약으로 인해 단일화된 BN의 복잡도는 더 증가할 가능

표 4 39개의 모듈형 베이지안 네트워크와 하나로 모델링 된 BN의 구조 정보. (MonoBN – 하나의 BN으로 설계된 monolithic Bayesian network, NN: 노드의 수, NNR: 루트 노드의 수, NNI : 중간 노드의 수, NNL: 리프(leaf) 노드의 수, NL: 링크의 수, NP_{avg}: 평균 부모의 수, NS: 상태의 수, NS_{avg}: 평균 상태의 수, NCPV: CPV의 수, CPT_{max}: CPT의 최대 크기)

	NN	NNR	NNI	NNL	NL	NPavg	NS	NSavg	NCPV	CPTmax
39 BNs	638	375	135	128	623	0.98	1,279	2.00	4,205	64
MonoBN	462	235	111	116	588	1.27	927	2.01	4.869	512

성이 높으므로 모듈화된 BN이 상대적으로 훨씬 유리하다.

4. 실험 및 결과

사용하는 로그 데이타는 GPS, Call, SMS, 사진 촬영, MP3 청취, 기기 충전 로그 및 웹에서 수집된 날씨 정보에서 추출된 로그 컨텍스트 데이타이다. 그리고 설계된 BN은 총 39개의 BN을 사용하였으며 실험의 객관성을 위해 실제 모바일 환경에서 테스트 하였다. 테스트에 사용된 기기는 HP의 Pocket PC인 iPAQ 1940이며, 운영체제는 Pocket PC 2003, 테스트를 위해 사용된 개발환경은 Microsoft eMbedded Visual C++ 4.00.1610.0이다.

4.1 사례 분석

제안하는 특이성 추론 모델의 성능을 관찰하기 위해 간단한 시나리오를 설정하고 실험을 수행하였다. 실험에서 사용된 25개의 증거 로그 컨텍스트는 {강의동, 레스토랑, 신촌, 자연지역, 집, 커피점, 평소가던곳, 학교지역, 학생회관, 2시간이내:외출, 낮, 식사시간, 아침식사시간, 아침식사전, 오전, 일광시간, 자기전2시간이내, 잘시간, 저녁식사시간, 저녁식사후, 점심식사시간, 디바이스사용없음, 사진많이찍음, 사진찍음, 야외활동성, 외출전, 이동중, 좋은날씨, 즐거움많은하루, 충전중, GPS잡힘}이다.

그림 8(a)는 실험에서 사용된 시나리오를 보여준다. 실험에 사용된 BN은 {외식, 사진, 이동행동, 자연, 즐거움, 집}의 6개이다.

그림 8(b)는 시나리오에 따라 생성된 하루 분량의 로그 컨텍스트를 이용해 특이성을 추론한 결과이다. 그림을 살펴보면 해당 시간에 관련된 특이성의 확률이 높아지는 것을 관찰할 수 있다. 예를 들어, 7~9시에는 '외출준비'와 '샤워' 특이성이, 12~13시와 17~19시에는 '식사' 특이성이, 13~14시와 20~21시에는 '산책' 특이성이, 14~15시에는 '즐거운 사진 찍기' 특이성이, 17~19시에는 '외식'과 '식사(서양)' 특이성의 확률이 높게 나타나고 있다. 그리고, 특이성을 살펴보면 하루 중 일부만 확률값이 나타나고 있는데, 이것은 관련 증거가 존재하지 않는 시간에는 해당 BN이 사용되지 않기 때문이다.

4.2 인공 데이타에 의한 성능 평가

본 특이성 추출기의 도메인은 매우 넓기 때문에 모든 상황을 고려한 평가는 어렵다. 따라서 특이성의 종류를 골고루 선택하여 평가하기 위해 일상생활에서 마주칠 수 있는 상황을 크게 4가지 상황(일상-한가함, 일상-바쁨, 비일상-한가함, 비일상-바쁨)으로 나누어서 각 상황을 대표하는 특이성을 생성하였다. 그림 9는 제시하는

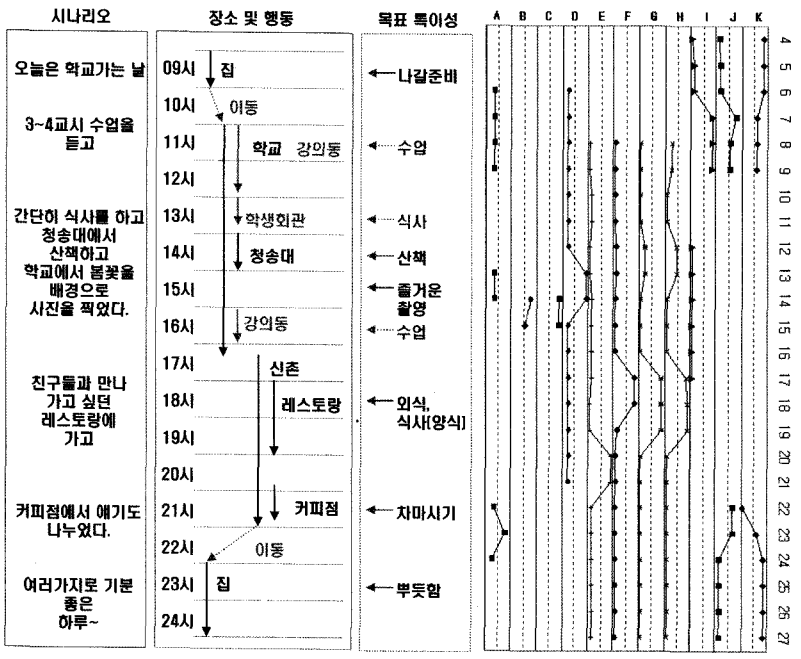

그림 8 (a) 실험을 위해 구성된 모바일 디바이스를 가진 대학생의 일상 생활 시나리오. (b) 11개의 목표 특이성에 대한 확률변화 관찰 결과. 표시된 숫자는 시각 4시~27시(다음날 3시)를 의미한다. 관련된 증거가 주어지면 확률값이 높아진다. (A: 뿌듯함, B: 사진 찍기(풍경), C: 즐거운 사진 찍기, D: 산책, E: 차 마시기, F: 외식, G: 식사(서양식), H: 식사, I: 외출 준비, J: 샤워, K: 수면)

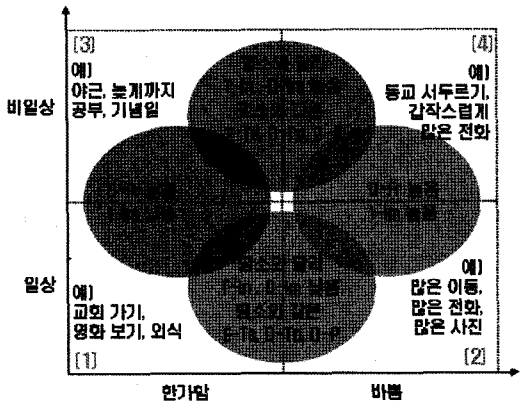

그림 9 특이성 분류 및 인공 데이타 특성 설명. 로그 데이타 분석 함수(T: time, D: daily, In: interval, Im: impact, Ts: time-span, Tp: time-portion, P: priority, Fr: frequency)를 기준으로 설명하였다.

기준으로 특이성 및 인공 데이타 특성을 분류한 도표를 보여준다.

베이지안 네트워크를 위한 전처리를 거치지 않은 인공 데이타를 직접 생성하는 것은 어렵기 때문에 그림 9에서의 4가지 구분을 따르되, 데이타는 특이성의 증거 컨텍스트를 기준으로 생성하였다. 예를 들어, 전화 통화의 Instant Frequency를 매 시간대 별로 생성하는 대신, Instant Frequency가 높은 시간대에 "전화통화많음" 컨텍스트를 증거로 정의하는 방법이다. 생성된 인공 데이타에 의해 추출되는 특이성은 '일상', '바쁨', '비일상'

표 5 일상, 바쁨, 비일상으로 나눈 특이성. 일상 특이성의 경우 일부만 표기하였다.

특이성 종류	특이성
일상 (81개)	공부중, 교통체증, 기차이용, 노래, 농구, 더움, 등산, 만남, 머리손질, 모임, 물건찍기, 미사, 뿌듯, 산책, 샤워, 선박이용, 설거지, 설렘, 세면, 셀카찍기, 쇼핑, 수면, 수상스키, 수영, … , 항공기이용, 행복, 헬스운동, 혼자공부, 화남, 화장, 황홀, 회식하기, 회의, 휴식
바쁨 (6개)	바쁨, 우산들고달리기, 즐거운통화, 즐거운SMS, 짜증SMS, 피곤한통화
비일상 (25개)	결혼, 경마관람, 공연관람, 관람, 기념일, 농구경기관람, 매우낯선곳, 비일상지역, 선거, 성묘, 야구경기관람, 여행, 영화관람, 입학식, 장례, 제사, 졸업식, 차례, 축구경기관람, 축하문자받기, 축하문자보내기, 축하전화걸기, 축하전화받기, 휴대폰분실, 휴대폰잊음

과 관련된 경우로 나누어 정의하였다. 앞의 네 가지 구분 항목에 따라 나누면 일상/한가={일상}, 일상/바쁨={일상}∪{바쁨}, 비일상/한가={일상}∪{비일상}, 비일상/바쁨={일상}∪{비일상}∪{바쁨}이다. 표 5는 '일상', '바쁨', '비일상'에 해당하는 특이성을 의미한다.

4가지 분류에 대해 각각 30일 분량의 인공 데이타를 생성하고 특이성 추출 성능을 평가하였다. 이때 각 종류별로 2개의 특이성이 포함되도록 하였으며,'비일상/바쁨'의 경우에는 '일상' 특이성 1개, '비일상'특이성 1개로 구성하였다. 표 6은 전체적인 실험 결과 통계를 보여주고 있으며, 표 7은 각 데이타에 대한 실험 결과를 보여준다. 이때, 기본 장소인 '집' 관련 특이성과 주요 특이성

표 6 실험 결과 통계. 각 종류 별로 2개의 목표 특이성을 랜덤으로 선택하여 인공 데이타를 생성하였다. (TP: true positive error rate (%), FP: false positive rate (%), FN: false negative rate (%))

종류	시간	목표 특이성의 수	TP	FP	FN	정확률	재현률
일상/한가함	30 days	60	46	14	14	0.767	0.767
비일상/한가함	30 days	58	43	10	15	0.811	0.741
일상/바쁨	30 days	55	41	2	14	0.953	0.745
비일상/바쁨	30 days	60	46	8	14	0.852	0.767
합계	120 days	233	176	34	57	0.838	0.755

표 7 '실험 결과 뽑힌 특이성 (데이타 타입 A 일부). 타입 B, C, D는 생략하였다. (OL: Number of Obtained Landmark, TP: True Positive, FP: False Positive, 데이타 타입: A (일상/한가), B (비일상/한가), C (일상/바쁨), D (비일상/바쁨))

데이타	목표 특이성	OL	TP	FP	뽑힌 주요 특이성
A-60301	체육대회, 기차이용	45	2	2	체육대회, 기차이용, 전철이동, 고속버스이용,
A-60302	선박이용, 기차이용	47	1	3	수영(실외), 전철이동, 고속버스이용, 기차이용,
A-60303	머리손질, 즐겁게음악듣기	19	1	0	머리손질, 화장,
A-60304	교통체증, 스노우보드	42	1	0	스키, 스노우보드,
…	…	…	…	…	…
A-60329	헬스운동, 수영(실내)	26	2	0	헬스운동, 수영(실내),
A-60330	항공기이용, 뿌듯	35	1	0	수영(실외), 항공기이용, 장거리
총합	60	866	46	14	

중요도 수치가 낮은 특이성은 주요 특이성에서 제외시켰다.

표 6의 전체적인 결과를 보면, '일상/한가'의 FP 오류가 높고 정확률이 낮은데, 이는 '일상' 특이성이 다양하고 다양한 장소에서 추출되기 때문에 상대적으로 구분이 어렵기 때문이다. 예를 들어, '선박이용' 특이성은 '바다' 장소에서 뽑히는데, '수영'특이성도 같이 잘못 뽑힌다. 특이성 종류가 '일상/바쁨'인 경우에는 전화를 많이 하거나 이동이 많은 등, 특징이 상대적으로 뚜렷하기에 정확률이 높았다. 재현률은 전체적으로 낮았는데, 모바일 환경의 로그만으로 인식하기 어려운 특이성이거나 BN 확률 테이블의 튜닝 부족으로 인한 것으로 보인다.

4.3 실제 수집 데이타에 의한 성능 평가

제안하는 특이성 추출기의 성능 평가를 위해 여대생이 실제 스마트폰을 가지고 27일 동안 일상생활 및 의무행동을 하면서 로그 데이타를 수집하였다. 입력된 컨텍스트의 수가 많고 적음은 GPS의 연결 상태에 크게 좌우되기 때문에 GPS가 수집되지 않은 날은 실험에서 제외하였다. 또한, 실험에서 GPS 데이타는 환경에 따라 자주 끊기고 실제 많은 활동을 했음에도 누락된 부분이 많았기 때문에 거짓 부정(False Negative) 오류에 대해서는 평가하지 않았다. 특이성의 선택 확률 기준치는 66%, 특이성 중요도 기준치는 5.0으로 두어 초과하는 결과에 대해서만 평가하였다. 실험에서 '일치', '부분 일치'를 판단하기 위해 사용자들이 매일 기록한 '활동 일지'와 실제 맵 상에서 나타나는 GPS 이동 정보가 이용되었다. 표 8은 실험 결과이다.

실험 결과를 살펴보면 완전 일치된 특이성(R_{HIT})은 평균 35% 정도로 낮았는데, 이는 수집된 데이타의 부정확성과 '활동 일지'의 내용만으로는 정확한 일치를 판단하기 어렵기 때문이다. 따라서, 본 실험에서는 여학생이 이동한 장소나 기록된 행동을 바탕으로 가능성이 있는 경우는 넓은 의미에서의 일치(R_{HIT}')까지 확대하여 평가하였다. 이 경우 일치율이 평균 89.4%로 상당히 높게 나왔다. 특히 학교에서 활동이 많은 경우 GPS 데이타에 의한 컨텍스트 정보가 충분하였기 때문에 일치율이 높게 나왔다. 표 9는 실제로 얻어진 특이성에 대해 '일치', '부분 일치'를 적용한 내용을 보여준다. 예를 들어 3월 2일의 경우 실험자는 동아리방에서 늦게까지 연습을 하였으나 해당 특이성이 존재하지 않으므로 '늦게까지공부' 특이성을 '부분 일치'로 간주하였다.

5. 결론 및 토의

본 논문에서는 모바일 디바이스 환경에서 작동하기 유리한 특이성 추론 모델을 제안하였다. 모바일 환경에서 효율적으로 작동할 수 있도록 모듈화된 구조를 제시하였으며, 모듈화된 상태에서의 유기적인 연결을 위해 가상 노드 개념을 이용해서 확률 증거 연결을 적용한 2단계 추론 방식을 소개하였다. '일상/비일상', '바쁨/한가함'을 기준으로 나누고 생성한 가상 데이타를 사용하여 실제 모바일 디바이스에서 수행한 실험에서는 의도했던 특이성이 잘 추출됨을 보였다. 실제 사용자에 의한 데이타 수집 및 평가도 소개하였다. 수집된 로그 데이타를 바탕으로 사용자가 작성한 일지와 추출된 특이성을 비교한 결과 비교적 높은 추출 성공률을 보였다.

하지만 소규모의 데이타에 대한 실험 결과이기 때문에, 향후 더 많은 실험 및 추론 모델의 개선이 필요할 것이다. 또한, 제안하는 가상 증거 기술은 1단계의 확률 증거는 모듈화되기 전과 같은 결과를 보일 수 있지만, 여러 단계를 거치거나 동시에 전달되는 증거의 경우에

표 8 실제 스마트폰을 이용하여 수집한 로그 데이타. '활동 일지'를 기록 하지 않은 날, GPS 데이타가 전혀 수집되지 않은 날, 서울시를 벗어나 활동한 날은 제외시킴. (N_{Con}: 입력 컨텍스트의 수, N_{LM}: 추출된 특이성의 수, N'_{LM}: 하루 동안 중복된 특이성과 가중치가 낮은 특이성 제외한 특이성의 수, N_{ERR}: 불일치 된 특이성의 수, R_{HIT}: N_{HIT}에 대한 일치율, R_{HIT}' : N_{HIT}'에 대한 일치율, R_{ERR}: N_{ERR}에 대한 에러율)

날짜	N_{Con}	N_{LM}	N_{LM}'	N_{HIT}	N_{HIT}'	N_{ERR}	R_{HIT}	R_{HIT}'	R_{ERR}
60224	116	72	13	3	10	0	23.1%	100.0%	0.0%
60227	167	49	15	4	11	0	26.7%	100.0%	0.0%
60228	64	50	8	3	4	1	37.5%	87.5%	12.5%
60302	202	128	18	8	10	0	44.4%	100.0%	0.0%
60304	102	53	7	1	5	1	14.3%	85.7%	14.3%
60306	86	56	12	5	3	4	41.7%	66.7%	33.3%
60308	114	92	12	3	7	2	25.0%	83.3%	16.7%
60309	103	45	7	4	2	1	57.1%	85.7%	14.3%
60315	128	76	13	4	9	0	30.8%	100.0%	0.0%
60317	46	45	8	3	3	2	37.5%	75.0%	25.0%
60321	67	40	10	4	4	2	40.0%	80.0%	20.0%
총11일	1195	706	123	42	68	13	34.1%	89.4%	10.6%

표 9 실험자의 GPS 이동 기록과 '활동 일지'를 바탕으로 일치 여부 판정 결과이다. 부분 일치는 '활동 일지'에는 없으나 방문 장소 및 행동을 기준으로 가능성이 높은 경우이다.

날짜	일치	부분 일치	불일치
60224	식사(한식), 모임, 쇼핑	공부중, 수업중, 수업시간, 바쁜시간, 짜증SMS, 관람, 외식, 교통체증, 노래방, 댄스장	-
60227	공부중, 즐거운사진찍기, 모임, 쇼핑	수업시간, 바쁜시간, 짜증SMS, 음식찍기, 물건찍기, 풍경찍기, 관람, 식사(한식), 외식, 노래방, 댄스장	-
60228	바쁜시간, 즐거운통화, 모임	수업시간, 짜증SMS, 관람, 공부중	쇼핑
60302	수업시간, 바쁜시간, 즐거운통화, 식사(한식), 외식, 모임, 산책	짜증SMS, 음식찍기, 물건찍기, 풍경찍기, 관람, 실망, 쇼핑, 노래방, 댄스장, 공부중, 늦게까지공부	-
60304	관람	컴퓨터작업, 세면, 바쁜시간, 짜증SMS, 교통체증	수업시간
60306	모임, 짜증SMS, 수업시간, 바쁜시간, 식사(한식)	산책, 관람, 교통체증	외식, 쇼핑, 노래방, 댄스장
60308	수업시간, 식사(한식), 외식	짜증SMS, 바쁜시간, 교통체증, 모임, 쇼핑, 노래방, 댄스장	관람, 차마시기
60309	공부중, 모임, 수업시간, 관람, 모임	바쁜시간, 교통체증	쇼핑
60315	수업시간, 식사(한식), 외식, 쇼핑	청소, 요리, 설거지, 바쁜시간, 관람, 모임, 노래방, 댄스장, 산책	-
60317	공부중, 산책, 수업시간	바쁜시간, 실망, 모임	쇼핑, 관람
60321	늦게까지공부, 수업시간, 식사(한식), 쇼핑	바쁜시간, 모임, 노래방, 댄스장	관람, 외식

는 같은 결과를 보이기 어렵다. 따라서 BN의 모듈화 및 가상증거기술에 대한 장단점 분석과 향상된 방법에 대한 연구가 필요하다. 그리고, 향후에는 좀더 넓은 실제 도메인에서 제안하는 방법을 장기적으로 적용한 뒤 이를 평가 및 검증하는 작업이 필요할 것이다.

참 고 문 헌

[1] M. Raento, A. Oulasvirta, R. Petit, and H. Toivonen, "ContextPhone: A prototyping platform for context-aware mobile applications," IEEE Pervasive Computing, vol. 4, no. 2, pp. 51-59, 2005.

[2] A. Oulasvirta, "Finding meaningful uses for context-aware technologies: The humanistic research strategy," Proc. Conf. Human Factors in Computing Systems, ACM Press, pp. 247-254, 2004.

[3] G.D. Abowd and E.D. Mynatt, "Charting past, present, and future research in ubiquitous computing," ACM Trans. Computer-Human Interaction, vol. 7, no. 1, pp. 29-58, 2000.

[4] A. Schmidt, A. Takaluoma, and J. Mntyjrvi, "Context-aware telephony over WAP," Personal Technologies, vol. 4, no. 4, pp. 225-229, 2000.

[5] Y. Nakanishi, T. Tsuji, M. Ohyama, and K. Hakozaki, "Context aware messaging service: A dynamical messaging delivery using location information and schedule information," Journal of Personal Technologies, vol. 4, no. 4, pp. 221-224, 2000.

[6] A. Krause, A. Smailagic, and D. P. Siewiorek, "Context-aware mobile computing: Learning context-dependent personal preferences from a wearable sensor array," IEEE Trans. on Mobile Computing, vol. 5, no. 2, pp. 113-127, 2006.

[7] R. DeVaul, M. Sung, J. Gips, and A. Pentland, "MIThril 2003: Applications and Architecture," Proc of 7th IEEE Int. Symposium on Wearable Computers, pp. 4-11, 2003.

[8] P. Zheng and L. M. Ni, "The rise of the smart phone," IEEE Distributed Systems Online, vol. 7, no. 3, 2006.

[9] P. Korpipaa, J. Mantyjarvi, J. Kela, H. Keranen, and E.-J. Malm, "Managing context information in mobile devices," IEEE Pervasive Computing, vol. 2, No. 3, pp. 42-51, 2003.

[10] J. Gemmell, L. Williams, K. Wood, R. Lueder, and G. Bell, "Pervasive capture and ensuing issues for a personal lifetime store," Proc. of the 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences, pp. 48-55, Oct. 2004.

[11] D.P. Siewiorek, A. Smailagic, J. Furakawa, A. Krause, N. Moraveji, K. Reiger, J. Shaffer, and F.L. Wong, "SenSay: A context-aware mobile phone," Proc. 7th Int. Symp. of Wearable Computers, pp. 248-249, Oct. 2003.

[12] P. Dourish, "What we talk about when we talk about context," Personal and Ubiquitous Computing, vol. 8, no. 1, pp. 19-30, 2004.

[13] K. B. Korb, and A. E. Nicholson, Bayesian Artificial Intelligence, Chapman & Hall/CRC, 2003.

[14] E. Horvitz, P. Koch, R. Sarin, J. Apacible, and M. Subramani, "Bayesphone: Context-sensitive policies for inquiry and action in mobile devices," Proc. of the Conf. on User Modeling, pp. 251-260, 2005.

[15] G. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," Machine Learning, vol. 9, pp. 309-347, 1992.

[16] D. Heckerman, "A tutorial on learning with Bayesian networks," Learning in Graphical Models, pp. 301-354, Dordrecht: Kluwer, 1998.
[17] E. Horvitz, S. Dumais, and P. Koch. "Learning predictive models of memory landmarks," CogSci 2004: 26th Annual Meeting of the Cognitive Science Society, pp. 1-6, 2004.
[18] P. Korpipaa and J. Mantyjarvi, "An ontology for mobile device sensor-based context awareness," Proc. Context'03, Lecture Note in Artificial Intelligence, no. 2680, pp. 451-459, Springer-Verlag, 2003.

황 금 성

2001년 2월 연세대학교 컴퓨터과학과 졸업(학사). 2003년 2월 연세대학교 컴퓨터과학과 졸업(석사). 2004년 3월~현재 연세대학교 컴퓨터과학과 박사과정 재학중. 관심분야는 진화 알고리즘, 지능형 에이전트, 베이지안 네트워크

조 성 배

1988년 연세대학교 전산과학과(학사) 1990년 한국과학기술원 전산학과(석사) 1993년 한국과학기술원 전산학과(박사) 1993년~1995년 일본 ATR 인간정보통신연구소 객원 연구원. 1998년 호주 Univ. of New South Wales 초청연구원. 1995년~현재 연세대학교 컴퓨터과학과 정교수. 관심분야는 신경망, 패턴인식, 지능정보처리

이 종 호

1988년 서울대학교 심리학과, 계산통계학과 부전공(학사). 1991년 서울대학교 심리학과(석사). 1995년 포항공과대학교 전산학과(석사). 2004년 서울대학교 인지과학협동과정(공학박사). 2004년~현재 삼성종합기술원 전문연구원. 관심분야는 웹 마이닝, 지식검색, 지능정보처리