

동적 베이지안 네트워크를 이용한 모바일 라이프로그 기반 사용자 행동 예측

박한샘*, 조성배*

*연세대학교 컴퓨터과학과

e-mail:sammy@sclab.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

Prediction of User Activity based on Mobile Life-log using Dynamic Bayesian Network

Han-Saem Park*, Sung-Bae Cho*

*Dept of Computer Science, Yonsei University

요 약

개인화 장비 기술의 발달과 함께 최근 모바일 디바이스는 카메라, MP3 플레이어 등 다양한 기능을 포함하고 있으며, 많은 사용자가 이를 사용하고 있다. 모바일 디바이스는 사용자가 항상 휴대하기 때문에 사용자 정보를 습득하기에 유용하며 따라서 이로부터 수집된 다양한 정보를 바탕으로 최근 여러 가지 서비스를 제공하기 위한 노력이 이루어지고 있다. 본 논문에서는 사용자의 모바일 로그를 바탕으로 행동 패턴을 파악하여 사용자가 앞으로 취할 행동을 예측하고자 하며, 이 과정에서 다양한 행동 패턴 중 정확한 행동 예측을 수행하기 위해 다음과 같은 방법을 활용하였다. 장소, 시간, 요일 정보를 함께 사용하여 동적 베이지안 네트워크를 이용해 시간 변화에 따른 사용자 행동 패턴을 학습하였으며, 개인 사용자 모델과 전체 사용자 모델을 따로 학습함으로써 더 정확한 행동 패턴의 학습이 가능하도록 하였다. 실험을 위해 대학생들로부터 수집된 모바일 로그를 통해 제안하는 행동 예측 모델의 성능을 확인한 결과 77~94%의 예측 정확도를 보임을 확인하였다.

1. 서론

최근 스마트폰, PDA와 같은 모바일 디바이스의 사용이 일반화 되었으며, 개인화 장비 기술의 발달로 카메라, MP3 플레이어 등 다양한 기능을 포함한 모바일 디바이스 또한 많이 사용되고 있다. 이러한 모바일 디바이스는 사용자가 항상 휴대하기 때문에 사용자의 개인 정보를 습득하기 위한 유용한 수단이 되는데, 이를 활용하여 사용자의 일상 생활을 저장, 관리하거나 사용자가 필요로 하는 서비스를 제공하고자 하는 다양한 시도가 이루어지고 있다[1].

가장 기본이 되는 모바일 로그정보는 GPS로부터 얻은 사용자 위치로, 이를 활용한 위치 기반 서비스(LBS, Location-based services)는 이미 많은 연구가 이루어지고 있으며[2, 3], 간단한 LBS를 기반으로 한 다양한 상용 서비스 또한 이루어지고 있다. 사용자 위치 외에도 PIMS(Personal Information Management System) 정보, MP3 음악정보, SMS, 사진 촬영 등 여러 가지 정보를 이용해 사용자의 행동이나 감정을 추론하는 연구가 수행되고 있다[1, 4, 5]. 또한 모바일 로그 정보를 추가적으로 분석하고 정리하여 사용자에게 흥미있는 형식으로 보여주기도 하는데, 사용자의 하루 일과를 만화 형식으로 보여주고자 한 AniDiary는 그 대표적인 예라고 할 수 있다[1].

본 논문은 모바일 라이프 로그 정보를 활용하여 사용자의 행동을 예측하고자 한다. 사용자의 모바일 로그로부

터 과거 행동 패턴을 분석할 수 있으며, 과거의 행동 패턴을 바탕으로 앞으로의 행동을 예측할 수 있다면, 사용자가 필요로 하는 정보를 미리 제공해 주는 등의 유용한 서비스가 가능하다. 사용자의 행동을 예측함에 있어서 가장 어려운 점은 한 사람의 가능한 행동의 범위가 너무 넓어서 정확한 예측이 불가능한 데에 있으며, 이러한 어려움으로 인해 사용자의 행동 예측은 거의 연구가 이루어지고 있지 않다. 다만 센서 정보를 활용해 사용자의 현재 행동을 인식하고자 하는 연구가 많이 수행되고 있으며[4, 5], 행동 수준의 예측은 아니지만 한정된 도메인에서 사용자의 이동 경로를 예측하고자 한 연구가 수행된 바 있다[6].

사용자의 행동을 예측하기 위해 본 논문은 사용자의 과거 행동 패턴을 동적 베이지안 네트워크를 이용해 학습하였다. 이 때 앞서 언급된 어려움을 극복하기 위해 행동과 함께 행동이 이루어진 장소, 시간, 요일 정보를 함께 사용함으로써 정확도를 높이고자 하였다. 또한 사용자 개인이 다른 행동 패턴을 가질 수 있으므로, 전체 사용자의 행동 예측 모델과 별도로 개인 사용자의 행동 패턴을 각각 학습하고, 전체 사용자 모델은 개인 사용자 모델을 보완하기 위해 사용하였다. 학습 과정에서 사용자의 행동과 장소를 나누는 기준이 중요한 역할을 하게 되는데, 객관적인 분류를 위해 관련 연구에서 비교적 최근에 수행된 행동 및 장소 분류 기준을 적용하였다[7, 8].

2. 행동 및 장소 분류

동적 베이지안 네트워크 모델을 학습하기 위해 모든 행동 및 모든 장소를 따로 분류할 수 없으므로 본 논문은 수집된 데이터의 행동 및 장소를 분류하기 위해 Statistics Canada에서 조사한 GSS (General Social Survey on Time Use)와 미국 University of Maryland, Survey Research Center에서 조사한 NHAPS (National Human Activity Pattern Survey)의 분류기준을 각각 활용하였다.

GSS는 캐나다에서 매 년 사회 변화를 파악하기 위해 수행되는 서베이로 여러 가지 주제에 대해 이루어지며 본 논문은 2006년 수행된 사람들의 시간 사용(time use)에 대한 조사에서 행동 분류 기준을 가져와 활용하였다[7]. GSS에 의하면 행동은 10가지 대분류, 24가지 중분류, 177가지 소분류로 나뉘며, 이 중 수집된 데이터에서 사용한 소분류는 모두 18가지이다. 10가지 대분류는 다음과 같다.

- Paid work and related activities
- Household work and relative activities
- Social support, civic and voluntary activity
- Education and related activities
- Socializing
- Television, reading and other passive leisure
- Sports, movies and other entertainment events
- Active leisure
- residual

NHAPS는 US Environmental Protection Agency의 후원으로 1992~1994년 조사한 데이터로 사람들의 하루 24시간 동안 행하는 행동 및 위치하는 장소 등의 패턴을 조사한 자료이다[8]. 본 논문에서는 장소 분류를 사용하였으며 사용된 8가지 장소 분류는 다음과 같다.

- Traveling/Near Vehicle (Outdoor)
- School/Church/Hospital/Public Building
- Residence (Indoor)

- Traveling Inside Vehicle
- Bar or Restaurant
- Mall/Grocery Store/Other Store
- Other Outdoor
- Other Indoor

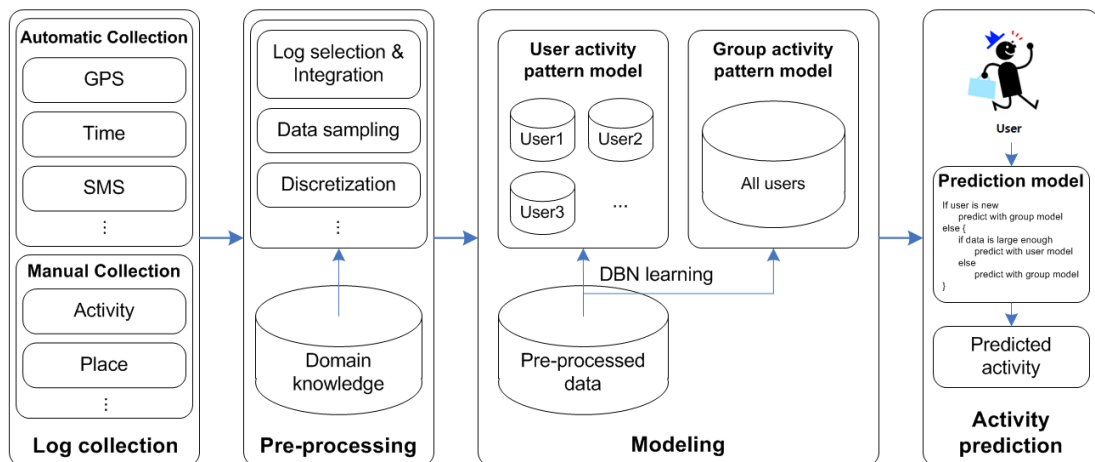
위와같은 기준으로 분류된 행동과 장소는 베이지안 네트워크의 행동과 장소 변수의 상태로 설정되므로 데이터를 분류하는 과정은 학습을 위한 전처리에 해당한다.

3. 제안하는 방법

본 논문에서는 사용자의 행동을 분류하기 위해 사용자의 모바일 기기에서 수집된 로그정보의 시퀀스를 입력으로 활용하였다. 기본적으로 사용자가 과거에 어떤 행동을 했었는지로부터 앞으로 어떤 행동을 할지를 예측하고자 하였으며, 과거의 행동 이외에 행동에 영향을 줄 수 있을 만한 요소인 장소, 시간, 요일 정보를 함께 활용하여 보다 정확한 행동 예측이 가능하도록 하였다. 그림 1은 제안하는 사용자 행동 예측 방법의 흐름도를 보여준다. 앞의 정보를 활용하여 행동 예측 모델을 획득하기 위해서 동적 베이지안 네트워크(Dynamic Bayesian Network, DBN)를 이용하였다.

3.1. 데이터 전처리

본 논문에서는 1분단위로 수집된 모바일 로그정보를 이용하며, 여기에 사용자가 직접 입력한 위치 등 부가정보가 일부 함께 활용된다. (데이터에 대한 설명은 다음 장의 실험 데이터 부분에서 상세히 기술된다.) 데이터 전처리 과정에서는 수집된 로그정보를 10분 단위로 다시 선택하였으며, 전날의 정보가 다음날 정보와 연결이 되지 않도록 데이터를 일별로 구분하였다. 이렇게 얻어진 하루의 데이터 가운데 스트림 형태로 5개, 10개, 15개, 20개, 25개의 로그가 하나의 시퀀스로 묶여 하나의 데이터가 되며, 가장



(그림 1) 제안하는 사용자 행동 예측 방법: 흐름도

오래된 데이터를 빼고 새로운 데이터를 추가하여 다음 데이터가 구성되는 방식을 반복하여 시퀀스가 만들어진다. 이렇게 구성된 시퀀스 데이터는 베이지안 네트워크의 학습을 위해 활용된다.

사용자의 행동, 장소, 시간, 요일의 분류 또한 데이터 전처리에 포함된다. 장소와 시간, 요일은 각각 사용자가 취하는 행동에 영향을 미칠 수 있으므로 이의 분류는 행동 예측 성능에 직접적으로 영향을 미칠 수 있다. 본 논문은 이러한 인과관계를 최대한 부각시킬 수 있는 기준을 찾고자 하였다. 행동 및 장소는 2장에 설명된 것과 같이 GSS와 NHAPS를 이용하여 각각 분류되었으며, 시간은 하루 24시간을 6시간 단위로 4등분하여 새벽(0시~6시), 아침(6시~12시), 오후(12시~18시), 밤(18시~24시)으로 분류되었다. 마지막으로 요일은 평일(월~금), 토요일, 일요일로 분류되었다.

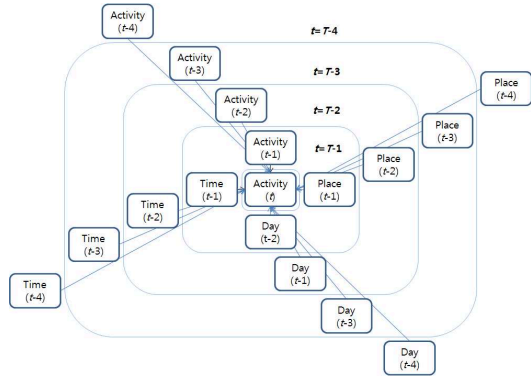
3.2. 동적 베이지안 네트워크를 이용한 예측 모델 학습

앞서 언급했듯이 본 논문에서는 과거의 장소, 시간, 요일, 행동의 시퀀스로부터 사용자가 다음에 취할 행동을 예측하기 위한 모델로 DBN을 사용하였다. DBN은 일련의 확률변수들의 조건부 확률 분포를 표현하기 위한 확률 모델로서 특히 변수들 간의 시간적인 관계를 표현하는데 유용한 모델이다[9]. 일반적으로 DBN 모델에서는 마르코프 특성에 의해 현재 시점의 확률 변수는 이전 N단계까지의 영향을 받고 그 이전의 시간단계에서는 영향을 받지 않는다. 따라서 추론 과정에서는 네트워크 구조에서 N단계만큼 unrolling 된 새로운 네트워크에 의해 각 확률 변수들의 조건부 확률 분포를 아래 식 (1)과 같이 계산해야 한다.

$$P(Z_1, \dots, Z_T) = \prod_{t=1}^T \prod_{i=1}^N P(Z_t^i | Pa(Z_t^i)) \quad (1)$$

사용자의 행동은 장소, 시간, 요일 등의 정보와 연관이 있으며, 사용자가 그 전에 취했던 행동과도 연관이 있을 수 있기 때문에 본 논문에서는 이러한 시간 변화에 따른 인과관계를 동적 베이지안 네트워크를 학습하여 모델링하였으며 이를 통해 사용자의 다음 행동을 예측하고자 하였다. 사용된 동적 베이지안 네트워크는 과거의 모든 증거 노드가 다음 행동 노드의 연결되는 naive Bayes 구조와 같은 모양을 취하며 데이터에 따라 얼마나 이전 단계까지의 증거를 사용할지를 달리 결정하였다. 파라미터 학습을 위해서는 MLE(Maximum Likelihood Estimation)가 사용되었다.

그림 2는 행동 예측을 위해 본 논문에서 사용된 동적 베이지안 네트워크의 가장 간단한 예이다. t=T에서의 행동을 예측하기 위해 t=T-4부터 t=T-1까지의 행동, 장소, 시간, 요일 정보를 함께 사용하여 베이지안 네트워크를 구성하였다.



(그림 2) 동적 베이지안 네트워크 예

3.3. 행동 예측

예측 모델의 학습은 사용자 별로 따로 수행되었으며, 이와 별도로 전체 사용자의 행동 예측 모델도 따로 학습되었다. 새로운 사용자가 예측을 요청하면 어느정도 데이터가 확보되어 안정적인 성능을 낼 수 있을 때까지 전체 사용자 모델을 이용해 예측을 수행하며, 일정 이상의 데이터가 확보된 사용자에 대해서는 개인 사용자 모델을 사용하여 행동을 예측한다.

4. 실험 및 결과

본 논문은 제안하는 방법에 의해 생성된 모델이 과거의 로그로부터 앞으로의 행동을 얼마나 정확히 예측하는지를 실제 사용자들에 의해 수집된 모바일 로그 데이터를 통해 확인하였다.

4.1. 실험 데이터

실험에 사용된 데이터는 5명의 대학생이 30일동안 스마트폰(삼성 m-4650)을 사용하며 수집한 로그 데이터를 사용하였으며 위치정보 수집을 위해서 추가적으로 휴대용 GPS인 BT-335를 함께 이용하였다. 수집한 정보는 스마트폰을 통해 자동으로 수집한 통화 내역, GPS, MP3 정보, 시간, 요일, 사진정보, SMS, 그리고 사용자가 레이블링을 해서 수동으로 수집한 행동과 장소가 있다. 본 논문에선 이 중 행동과 밀접한 관계가 있다고 판단되는 행동, 장소, 시간, 요일 정보를 사용하였다.

4.2. 행동 예측 실험 결과

3.3장에서 설명했듯이 본 논문에서는 개인 사용자 모델과 전체 사용자 모델을 따로 학습하였다. 표 1은 사용자 각각의 행동 예측 모델을 이용한 예측 실험 결과(정확도)를 보여주며, 표 2는 전체 사용자 모델을 이용한 예측 실험 결과를 보여준다. 2장에서 설명된 행동 중 실제로 수집된 행동은 모두 18가지이며, 사용자 1 ~ 5는 각각 4가지, 3가지, 6가지, 4가지, 4가지의 행동을 하였다.

〈표 1〉 개인 사용자 모델의 행동 예측정확도(%)

시퀀스 수 \ 예측 모델	5	10	15	20	25
사용자 모델 1	94.02	75.87	64.70	69.42	69.77
사용자 모델 2	82.58	75.26	76.83	77.33	80.96
사용자 모델 3	56.64	58.03	59.49	70.23	68.50
사용자 모델 4	77.78	75.07	75.28	63.95	-
사용자 모델 5	78.44	78.48	74.22	74.23	74.63

〈표 2〉 전체 사용자 모델의 행동 예측정확도(%)

시퀀스 수	5	10	15	20	25
예측정확도(%)	47.73	42.01	43.86	47.93	49.44

시퀀스 수는 DBN에서 고려한 시점의 수(5인 경우 앞의 4시점까지를 고려)를 의미한다. 전체적으로 적은 수의 시퀀스라도 정확한 예측이 가능한 것처럼 보이지만 항상 그렇지는 않으며, 전체 사용자 모델의 경우도 25개의 시퀀스 수를 사용했을 때의 정확도가 가장 높다. 사용자 별로 차이를 보이는 것은 사용자별로 행동의 수가 다르고 행동 패턴(혹은 장소, 시간, 요일과의 관계 패턴)이 얼마나 잘 나타나는지의 차이에 의한 것이다. 표 6에 나타난 전체 사용자 모델의 경우 40%대의 낮은 예측율을 보이지만 전체 행동수가 18가지라는 점을 감안하면 그리 낮은 수치는 아니다. 개인 사용자 모델의 예측율은 전체 사용자 모델의 예측율보다 훨씬 높은 77~94%의 꽤 정확한 예측율을 보임을 확인할 수 있다. 사용자 모델 4에서 25개의 시퀀스를 사용하는 경우는 사용자가 기록한 행동의 수와 비교해 시퀀스 수가 너무 커서 전처리 과정을 거친 후 학습을 할 수 있는 충분한 데이터가 확보되지 않아 실험을 하지 못하였다.

5. 결론 및 향후연구

본 논문은 모바일 디바이스로부터 수집된 로그를 이용하여 사용자의 향후 행동을 예측하는 문제를 풀고자 하였다. 행동 예측을 위해 행동과 관련성이 깊은 정보인 장소, 시간, 요일 정보를 함께 이용하였으며, 과거의 로그를 스트림 형태로 입력받아 향후 행동을 예측하기 위해 사용하였다. 행동 분류를 위해서는 GSS의 행동 분류 기준을 사용하였으며, 장소 분류를 위해서는 NHAPS의 장소 분류 기준을 사용하였다. 이와 같은 전처리 과정을 거친 후 행동 예측 모델을 학습하기 위해 동적 베이저안 네트워크를 이용하였다. 이 때, 개인 사용자 모델과 전체 사용자 모델을 따로 학습하였으며, 새로운 사용자에 대해서 데이터가 일정 이상이 확보되기 전까지는 전체 사용자 모델을 이용해 행동 예측을 수행할 수 있도록 하였다. 5명의 사용자로부터 수집된 모바일 로그를 이용해 만들어진 모델의 예측 정확도를 확인해 보았으며 전체 사용자 모델의 경우 49%, 개인 사용자 모델의 경우 77~94%의 예측 정확도를 보여

그 가능성을 확인하였다.

본 논문에서는 행동 및 장소, 시간, 요일 로그를 이용해 향후 행동을 예측하였다. 하지만 모바일 로그는 통화 내역, SMS, MP3, 사진 촬영 등 더 많은 정보를 포함한다. 이러한 추가정보를 잘 활용하면 더 높은 예측 정확도를 갖는 모델을 학습할 수 있을 것이다. 예를 들면, 잦은 통화나 SMS의 잦은 사용은 특정 행동과 연관성을 가질 수 있다. 일단 보다 정확한 행동 예측이 이루어지면, 이를 바탕으로 사용자가 향후 필요로 하는 정보를 추천해 주는 등의 유용한 서비스 제공 또한 가능할 것이다.

감사의 글

이 논문은 한국과학기술연구원 (KIST) 실감형 차세대 웹 기술개발 과제에 지원을 받아 수행된 연구임.

참고문헌

- [1] S.-B. Cho, K.-J. Kim, K.-S. Hwang and I.-J. Song, "AniDiary: Daily cartoon-style diary exploits Bayesian networks," *IEEE Pervasive Computing*, pp. 66-75, 2007.
- [2] P. Bellavista, A. Kupper and S. Helal, "location-based services: Back to the future," *IEEE Pervasive Computing*, vol. 7, no. 2, pp. 85-89, 2008.
- [3] J. Krumm and E. Horvitz, "Predestination: Where do you want to go today?" *Computer*, vol. 40, no. 4, pp. 105-107, 2007.
- [4] S.-W. Lee and K. Mase, "Activity and location recognition using wearable sensors," *IEEE Pervasive Computing*, vol. 1, pp. 24-32, 2002.
- [5] M. Ermes, J. Parkka, J. Mantjarvi and I. Korhonen, "Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 1, pp. 20-26, 2008.
- [6] S.-J. Han and S.-B. Cho, "Learning trajectory information with neural networks and the Markov model to develop intelligent location-based services," *Journal of Information and Knowledge Management*, vol. 5, no. 4, pp. 1-11, 2006.
- [7] Statistics Canada, General social survey on time use, <http://www.statcan.ca/>
- [8] N. E. Klepeis, et al., "The national human activity pattern survey (NHAPS): A resource for assessing exposure to environmental pollutants," *Journal of Exposure Science and Environmental Epidemiology*, vol. 11, pp. 231-252, 2001.
- [9] K. Murphy, "Dynamic Bayesian networks: Representation, inference and learning," Ph.D. Thesis, University of California, Berkely, 2002.