

과거 상황 정보의 변동 양상을 활용한 인간의 중단 가능성 예측 모델 연구

최민수 강민서 김민석 김선동

한국과학기술원 지식서비스공학대학원

{minsoo.choy, [minseo](#), minseokkim, sundong.kim} @kaist.ac.kr

Exploiting Change Patterns of the Past Context Data on Human Interruptibility Prediction

Minsoo Choy Minseo Kang Minseok Kim Sundong Kim

Graduate School of Knowledge Service Engineering, KAIST

요 약

우리는 일상을 살아가면서 다른 사람들과의 관계에 있어, 전화를 하거나 문자를 보내는 등 그들의 시간을 잠시나마 요하는 경우가 많다. 하지만, 모든 경우에 그들이 우리를 위해 시간을 내어 줄 수는 없다. 그렇기 때문에, 타인의 응답 가능성 (혹은 중단 가능성)을 최대한 정확하게 예측하는 연구가 최근 각광받고 있다. 이를 위해서는, 사람들이 현재 어떠한 상황에 있는지 혹은 어떠한 행동을 하고 있는지 파악할 수 있어야 하는데, 오늘날 우리 사회에 많이 보급되어 있는 스마트 기기들은 이를 가능하게 한다. 대다수의 많은 연구들은 사람의 상황을 파악할 때, 예측 시점에서의 상황정보만을 고려하는 경향이 있다. 하지만, 우리는 긴 과거의 행동도 현재의 중단 가능성에 영향을 줄 수 있다는 가정 하에, 보다 긴 과거의 정보들을 적극적으로 활용한다. 즉, 배터리 잔량이나 가속도 센서 등의 스마트폰 데이터들이 과거에 어떤 양상으로 변화 하였는지를 추가적으로 중단 가능성의 예측에 활용하였으며, 이러한 시도가 성능향상에 도움이 될 수 있음을 확인하였다.

1. 서론

현대인들은 모두가 바쁜 생활을 살아가고 있으며, 그 속에서 서로 간에 전화나 문자 메시지를 주고받는 등 끊임없는 의사소통이 이루어진다. 하지만, 이러한 의사소통이 매번 성사되는 것은 아니다. 수신자 측이 언제나 수신 가능한 상황에 놓여있는 것은 아니기 때문이다. 이러한 경우, 발신자 측은 상대방의 응답을 기다리기까지 계속 대기해야하는 상황이 발생한다. 이러한 이유로, 오늘날 수신자 측의 중단 가능성 (Interruptibility)을 예측해내는 연구가 각광받고 있다. 성공적인 중단 가능성 예측 모델은 발신자와 수신자 측 모두에게 큰 도움이 될 수 있다 [1].

중단 가능성의 연구를 정확하게 수행하기 위해서는 사용자들이 어떠한 상황에 놓여있는지 (Context Awareness)를 제대로 파악해내는 것이 매우 중요하다. 오늘날 많은 사람들은 스마트폰과 같은 스마트 기기들을 사용하고 있다. 이러한 기기 속에는 가속도 센서, 조도 센서를 비롯한 다양한 센서들이 내장되어 있기 때문에, 이러한 작업을 보다 쉽게 할 수 있다. 이런 스마트폰 상황 정보 (Context Data)에 기반한 예측은 매우 다각적으로 이루어지고 있다. 본 논문과

유사하게는 현재 문자메시지를 읽을 수 있는지의 여부 [2], 전화를 받을 수 있는지의 여부 [3]에서부터, 현재 사용할 만한 애플리케이션의 목록 예측 [4, 5] 등 매우 다양하다. 해당 연구들은 대부분 예측하는 시점의 상황 정보에 기반하여 모델링을 해낸다. 다시 말해서, 예측 시점에서의 배터리 잔량, 혹은 사용중인 애플리케이션 등이 Feature로서 예측에 사용되고 있는 것이다.

본 연구가 기존 연구들에 비해가지는 가장 큰 차이점은 현재에 가까운 정보뿐만 아니라, 긴 과거의 정보를 추가적으로 사용하는 것이다. 이는 과거의 행동 역시 현재의 중단 가능성에 어느정도 영향을 미칠 수 있을 것이라는 전제에 기반한다. 우리는 이를 위해, 과거에 스마트폰 데이터들이 어떤 양상으로 변화 했는지의 정보를 사용하며, 체계적인 연구를 위하여, 시간대 (아침, 점심 등) 별로 해당 데이터를 추출해낸다. 이런 정보들이 중단 가능성 연구에 유용하게 활용될 수 있기를 기대한다.

2. 중단 가능성 데이터 셋

우리는 본 연구를 위하여 Device Analyzer 데이터 셋을 사용한다. 본 데이터 셋은 현존하는 스마트폰 데이터 셋들 중에서 규모가 가장 큰 것으로 알려져 있으며, 각종 센서 데이터들과 전화 및 메시지 기록 등이 저장되어 있다. 우리는 비교적 오랜 기간 동안 데이터가 수집된 50명의 데이터를 사용하였으며,

본 연구는 '국토교통부 국토공간정보연구사업 국토공간정보의 빅데이터 관리, 분석 및 서비스 플랫폼 기술개발 (17NSIP-B081011-04) 과제'의 연구비 지원에 의해 연구되었음

예측하고자 하는 클래스 변수 (Interruptibility)는 전화 수신 여부에 따라 결정하였는데, ‘부재중 전화’ 혹은 ‘10초 이내로 끊은 경우’는 전화를 정상적으로 받을 수 없는 상황이라 간주하여 ‘Not Interruptible’로, ‘10초보다 긴 통화를 한 경우’는 ‘Interruptible’로 판단하였다. 우리가 본 데이터 셋으로부터 사용하는 데이터들은 표 1과 같다.

표 1. 예측에 사용되는 데이터 종류

| 이름 | 설명 | 타입 |
|------------------|----------------|---------|
| bat_level | 배터리 잔량 | Numeric |
| bat_temperature | 배터리 온도 | Numeric |
| accel | 가속도 센서 값 | Numeric |
| light | 조도 센서 값 | Numeric |
| vol_system | 시스템 볼륨 크기 | Numeric |
| vol_music | 음악 볼륨 크기 | Numeric |
| vol_notification | 알림 볼륨 크기 | Numeric |
| airplane | 비행기 모드 여부 | Binary |
| screen | 스크린 점등 여부 | Binary |
| headset | 이어폰 꽂힘 여부 | Binary |
| wifi_mode | 와이파이 설정 여부 | Binary |
| wifi_conn | 와이파이 연결 여부 | Binary |
| mobile_conn | 셀룰러 연결 여부 | Binary |
| bluetooth | 블루투스 설정 여부 | Binary |
| charge | 충전 여부 | Binary |
| ringtone | 일반/진동/무음 | Nominal |
| charge_stat | AC/USB/충전 중 아님 | Nominal |
| display_orient | 가로/세로/정방 | Nominal |
| app_pkg | App 패키지 이름 | Nominal |
| app_cat | App 카테고리 이름 | Nominal |
| location | 장소 | Nominal |
| Interrupt (클래스) | 중단 가능성 여부 | Binary |

3. 중단 가능성 모델

본 장에서는 중단 가능성 예측 모델의 전반적인 개요 및 과거의 데이터 변화 양상을 어떻게 Feature 값으로 활용하는지 간략히 설명한다.

3.1. 중단 가능성 모델 개요

모델의 기본적인 흐름은 그림 1과 같다. 사용자의 Interruptibility 데이터와 해당 시점에 상응하는 상황 정보 및 과거의 상황 정보를 이용하여 Random Forest 기법을 사용하여 모델링을 한다. 학습된 모델은 추후에 상황이 주어지면, 그 상황에 대한 사용자의 중단 가능성을 예측하는데 사용된다.

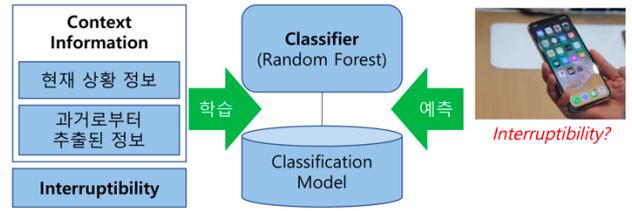


그림 1. 중단 가능성 모델의 개요

3.2. 과거로부터의 Feature 추출 방법

우선, 하루를 그림 2와 같이 세 시간 길이의 여섯 개 시간대로 나눈다.



그림 2. 여섯 개의 시간대

과거 정보는 표 1의 7개 Numeric 데이터에 대해서 추출한다. 우리는 각 Numeric 데이터마다, 각 시간대에 대해서 세 시간 길이의 시계열 (Time Series) 데이터를 구할 수 있다. 각 시계열은 증가경향이나 아래로 볼록 등 각자의 모양이 있다. 예를 들어, 아침에 배터리를 충전시켰다면, 해당 시계열은 증가하는 경향의 모양일 것이다. 우리는 각 Numeric 데이터마다 시간대별로 z-normalization된 시계열들의 집합을 구한 후, 각 집합에 대해 시계열 모양 기반 Agglomerative Clustering (Average Linkage)을 수행한다, 시계열 사이의 거리는 DTW (Dynamic Time Warping) distance를 사용한다. 그림 3은 한 사용자의 오후 시간대의 배터리 잔량 시계열 클러스터의 예시이다. 왼쪽 그림 (클러스터 ID #1)은 점점 감소하는 경향, 오른쪽 그림 (클러스터 ID #4)은 점점 증가하는 경향의 시계열들이 모여있음을 알 수 있다.

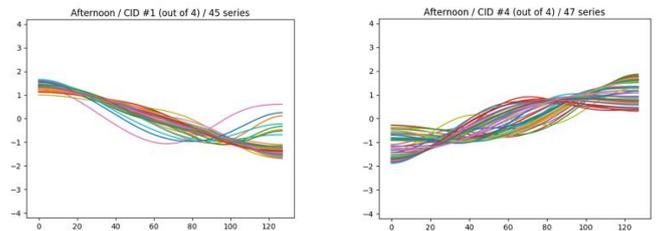


그림 3. 시계열 클러스터 예시

각 인스턴스의 시점 (전화 수신 시점)마다 해당 인스턴스가 소속된 시간대의 이전 시간대 각각에 대해서, 시계열이 소속된 클러스터의 ID를 Feature로 사용한다. 예를 들어, 점심의 인스턴스에 대해서는 ‘이른 아침’과 ‘아침’의 시계열이 속한 클러스터의 ID가 Feature로 쓰이게 된다. 즉, 이런 새로운 Feature들을 사용함으로써, 이전 시간대에서의 유저의 행동특성들을 어느 정도 예측에 활용할 수 있게 된다.

4. 실험

본 장에서는 과거로부터 추출된 Feature의 유용성을 평가하기 위한 실험 및 결과에 대해 설명한다.

4.1 전처리 과정

우리는 보다 좋은 모델링을 위하여, 크게 두 가지 전처리 과정을 가진다. 우선, 배터리 잔량같은 Numeric Feature 들에 대해서는 값을 그대로 사용하지 않고, MLDPCC를 이용해서 Discretization 시킨다. 그 후, 각 Feature들의 클래스와의 Correlation 에 기반해서, 상위 30개의 Feature만을 사용한다.

4.2 실험 및 결과에 대한 고찰

모델의 예측력에 대한 평가를 위하여, 우리는 평가지표로서 전체 예측 중 올바른 예측의 비율인 Accuracy를 사용한다. 평가 방법으로는 데이터 셋을 8:2로 나눈 후, 앞의 80%는 Training 셋으로, 뒤의 20%는 Test 셋으로 사용한다. 모델링은 예측시간대별로, 사용자 각각에 대한 개인화된 모델을 구축한다. 다시 말해, 사용자 각각에 대해, 아침부터 밤까지 각 시간대에 대한 모델을 구축한다. 이른 아침은 다른 시간대들에 비해 인스턴스의 개수가 드물어서 모델링에 고려하지 않았다. 우리는 현재 상황정보만을 이용하는 설정과 과거 정보를 함께 고려하는 설정 두가지를 비교하기로 하였고, 앞의 설정을 CURR (Current), 뒤의 설정을 PAST라 부르기로 한다. 표 2는 시간대 별 평균 Accuracy를 나타낸 것이다.

표 2. 시간대 별, 평균 중단 가능성 예측 정확도 (%)

| 아침 | | |
|------|------|-----------|
| CURR | PAST | PAST-CURR |
| 68.5 | 71.1 | +2.6 |
| 점심 | | |
| CURR | PAST | PAST-CURR |
| 69.5 | 72.9 | +3.4 |
| 오후 | | |
| CURR | PAST | PAST-CURR |
| 68.7 | 72.4 | +3.7 |
| 저녁 | | |
| CURR | PAST | PAST-CURR |
| 68.9 | 72.8 | +3.9 |
| 밤 | | |
| CURR | PAST | PAST-CURR |
| 67.3 | 70.3 | +3.0 |

표 2에서 볼 수 있듯이, 모든 예측 시간대에 대해서, PAST가 CURR에 비해 높은 정확도를 달성할 수 있음을 알 수 있다. 또한, 흥미로운 점은 아침부터 저녁까지는 뒤 시간대일수록, CURR 대비 PAST의 정확도 향상폭이

점점 커지는 사실이 발견되었다. 아마, 뒤 시간대일수록, 참고할 수 있는 앞 시간대에서의 행동 특성이 늘어나기 때문일 것이라 예상된다. 이로써, 과거 행동특성에 대한 고려는 중단 가능성 예측에 있어서 긍정적인 효과가 있음을 확인할 수 있었다.

5. 결론 및 향후 연구

본 논문에서는 인간의 중단 가능성을 예측하는데 있어, 예측 시점에서의 상황 정보만을 주로 고려하는 기존의 연구 방식과 달리, 오래된 과거의 정보를 보다 적극적으로 고려하는 시도를 하였다. 과거를 체계적이고 효과적으로 고려하기 위해서, 하루를 세 시간 길이의 시간대들로 나누고, Numeric 데이터들이 각 시간대에서 어떤 모양으로 변화 하였는지를 Feature로 사용하였고, 이를 위해 시계열 모양 기반 클러스터링을 활용하였다. 그 결과, 과거의 행동 특성이 현재의 상황정보만을 활용하는 것에 비해 중단 가능성 예측의 정확도가 많게는 3.9% 가량 올라가는 것을 확인할 수 있었다.

본 논문에서는 Numeric 데이터에 대해서만 과거를 고려하였으나, 향후 연구에서는 Binary 및 Nominal한 데이터에 대해서도 과거를 고려하여 보다 정확도를 향상시킬 계획이다. 예를 들어, 스크린 점등과 같은 Binary 데이터에 대해서는 세 시간 동안 몇 번이나 (혹은 몇 분 동안) 스크린이 켜져 있었는지 등의 정보를 활용할 수 있을 것이라 보이며, 이러한 Feature들 또한, 사용자의 과거 시간대에서의 행동특성을 효과적으로 반영할 수 있을 것이라 예상된다.

참 고 문 헌

[1] Turner, L. D. et al., "Interruptibility Prediction for Ubiquitous Systems: Conventions and New Directions from a Growing Field", in Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 801-812, 2015.

[2] Pielot M. et al., "Didn't You See My Message? Predicting Attentiveness to Mobile Instant Messages", In Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems, pp. 3319-3328, 2014.

[3] Pielot M., "Large-scale Evaluation of Call Availability Prediction", In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 933-937, 2014.

[4] Chin C. et al., "Understanding and Prediction of Mobile Application Usage for Smart Phones", In Proceedings of the 2012 ACM Conference on the Ubiquitous Computing, pp. 173-182, 2012.

[5] Huang K. et al., "Predicting Mobile Application Usage Using Contextual Information", In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 1059-1065, 2012.