와이파이 모니터링 기술로 센싱된 고객들의 매장 내부 이동 패턴들을 이용한 재방문 예측

김선동, 이재길

한국과학기술원 산업 및 시스템공학과

{sundong.kim, jaegil@kaist.ac.kr}

Predicting customer's revisit intention using indoor movements in stores by Wi-Fi monitoring

Sundong Kim, Jae-Gil Lee

Korea Advance Institute of Science and Technology

요 약

오프라인 매장에서 기존 고객의 재방문은 신규 고객 유치보다 중요한 요소로 작용한다. 따라서 매장의 관점에서는 재방문 의사가 높은 고객을 확인하는 방법을 필요로 한다. 이 논문에서는 와이파이 모니터링 기술로 센싱된 고객의 매장 내부 이동 패턴을 기반으로 오프라인 매장에서의 재방문을 결정하는 다양한 요소를 탐구하고 재방문율을 예측하는데 필요한 전처리 방법과 특성 추출 기법, 기계 학습 모델을 제시한다. 또한 오프라인 의류 매장 두 곳의 1년간 데이터를 통하여 매장 내부의 고객별 동선 데이터만으로도 안정적으로 60% 중반의 정확도를 보이는 재방문율 예측 분류기를 만들 수 있음을 확인하였다.

1. 서론

전자 상거래가 활성화된 시대이지만, 여전히 소매 시장 매출의 92%에 달하는 규모가 오프라인 매장에서 일어나고 있다[1]. 유저들이 웹사이트를 통해 구경한 제품들의 목록과 실제 구매한 제품을 로그로 쉽게 남길 수 있는 인터넷 쇼핑몰과 달리, 오프라인 쇼핑몰의 경우에는 유저들의 행적을 기록하기가 쉽지 않고, 이를 분석하여 고객의 재방문(revisit)을 이끄는 경우는 아직 기초 단계에 불과하다[2,3,4]. 본 연구에서는 고객들이 매장의 구역별 와이파이 센서 주위에 머무른 데이터를 이용해, 매장 내부에서의 동선(trajectory)을 추출하고, 이를 바탕으로 움직임 패턴에 관련된 다양한 특성들을 추출하여 고객들의 재방문 여부(revisit intention)를 예측하는 모델을 제안하였고, 두 곳의 오프라인 의류 매장의 실제 데이터를 통한 기계 학습으로 63% 성능의 분류기(classifier)를 만들었다. 본 논문의 구성은 다음과 같다. 2장에는 고객의 동선을 기준으로 하는 데이터 전처리 방법을 소개하였고, 3장에는 기계 학습 모델에 이용될 네 가지 분류의 특징 추출(feature engineering) 방법들을 서술하였다. 4장에는 두 오프라인 매장[1]의 데이터 통계와 기계학습 모델을 활용하여 얻어진 재방문율 예측 결과를 서술하였고, 5-6장에는 연구 발전 방향과 결론을 정리하였다.

2. 데이터 전처리

Wi-Fi AP(Access Point)를 통해 수집되는 매장 내 데이터 로그의 속성은 다음과 같다.

표 1. AP를 통해 수집되는 데이터 로그의 예시

속성	설명	예
device_id	암호화된 모바일 기기의 맥 어드레스	aafc10f061882da0e1a043
dwell_time	각 구역에서 머무른 시간	35
area	매장 내의 구역명	1f-c
deny	장기체류 고객이나 자주 방문한 고객	True
ts	시간 (Unix timestamp)	1449457815

표 1과 같은 속성을 지닌 로그들이 시간 순에 따라 서버에 저장되는데, 매장 내부 이동 패턴을 기준으로 하기 위해, 날짜와 device_id를 합친 정보를 기본 키(primary key)로 삼아 여러 개의 데이터 로그들을 시간 순으로 재결합하였다.

표 2. 움직임을 기준으로 재결합한 데이터 로그의 예시: device_id 'afc10f061882da0e1a043' 에 대한 방문 구역[2] (연결 시작 시간 오름차순 정렬)

[1] 본 논문의 실험에 쓰인 두 매장의 상호명 및 매장 내 구역명은 익명성을 유지하기로 했다.

[2] 구역 out은 매장 주변 유동인구의 범주에 들어온 시간을, 구역 in은 매장 내부로 인식되는 센서 범위 내에 들어온 시간을 의미

구역명	연결 시작 시간	연결 종료 시간
out	10:55:27	11:21:03
in	10:55:40	11:19:25
1f-a	10:55:41	10:57:30
1f-b	10:56:27	10:59:27
1f-c	10:58:10	11:01:45
2f-b	11:02:05	11:08:40
3f	11:13:02	11:15:56
1f-c	11:16:34	11:19:20

표 2와 같이 재결합된 데이터는 고객이 하루 동안 매장의 Wi-Fi AP에 센싱된 정보로 본 논문에서는 이동 패턴(moving pattern)이라고 칭한다. 또한 매장 내부로 인식되는 로그가 적어도 하나 이상 존재하는 이동 패턴을 내부 이동 패턴(indoor moving pattern)이라 칭한다. 같은 device_id라 하더라도, 방문하는 날짜에 따라 내부 이동 패턴은 다른 특징을 갖게 되며, 이 연구에서는 재방문을 예측하기 위해 내부 이동 패턴으로부터 다양한 특징 추출을 진행한다. 특징 추출 기법을 설명하기 이전에 이 연구에서 예측하고자 하는 재방문에 대해 정의하자.

정의: 특정 내부 이동 패턴이 센싱된 device_id에 대해, 재방문 간격(= 90일) 내에 다른 내부 이동 패턴이 감지되면, 해당 내부 이동 패턴은 재방문 의도는 True라고 할 수 있다.

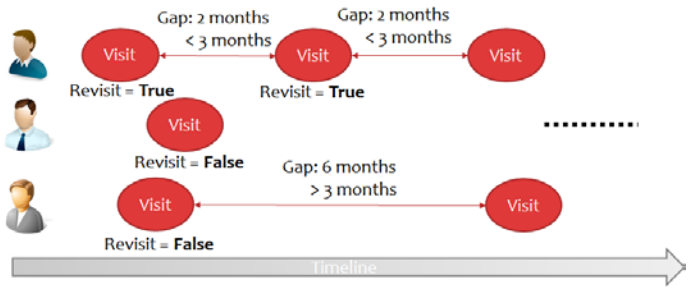

그림 1. 3명의 고객으로부터 발생한 6개의 내부 이동 패턴 중 체크 가능한 4개의 내부 이동 패턴에 대한 재방문 의도 여부

의류를 구매하는 고객의 경우 계절별로 매장에 방문한다는 점에 착안하여 재방문 간격을 90일(3달)로 설정하였다. 고객의 마지막 방문에 한해서는, 추후 90일 내에 재방문이 이루어질 지에 대한 여부를 알 수 없으므로, 최근 90일 내의 방문 중에서 고객의 마지막 방문의 경우에는 학습 및 검증 데이터에서 삭제해 준다. 그 예로 그림 1에서의 마지막 두 방문의 경우는 위와 같은 이유로 인해 재방문 의도를 파악할 수 없기 때문에 앞서 발생한 네 개의 방문과 그에 따른 재방문 의도 라벨만을 학습 및 검증 데이터셋으로 사용한다.

3. 특성 추출 방법

기계 학습을 위해 크게 네 가지 분류의 특성(feature)을 추출하였다. 먼저, 해당 내부 이동 패턴이 일어나기까지 고객이 방문한 총 횟수를 히스토리 데이터로 이용하였고, 장기 체류 구역(stay point)을 이용한 로그의 통계치와 방문이 일어난 요일, 마지막으로 동선의 n-gram(n=1,2) 특성값을 동선 스케쥴로부터 추출하였다. 두 번째 분류인 로그의 통계치의 경우, 해당 내부 이동 패턴에서 Wi-Fi에 연결된 구역의 총 개수, 총 시간, 머무른 시간이 100초 이상인 내부 장기 체류 구역 개수, 센싱된 내부 구역 개수 중 체류 시간이 100초 이상일 확률, 내부 장기 체류 구역에서 머무른 시간의 합과 표준편차 총 6가지 특성을 사용하였다. 요일의 경우 categorical variable이기 때문에 one-hot encoding을 통하여 7개의 이진 변수로 나타내 특성으로 사용하였다. 동선의 uni-gram, bi-gram의 예로는 '1f-a', '1f-b, 1f-a'가 있으며 해당 내부 이동 패턴이 n-gram을 몇 번 가지는지를 계산하여 총 123개의 특성으로 사용하였다. 도합 137개의 특성이 내부 이동 패턴 데이터로부터 생성되었고, 표 3와 같은 형태로 정리하여 기계 학습 모델에 이용하였다.

표 3. Feature engineering 이후 데이터 형태 예시

Key	F_1	F_2	F_3	F_4	.	F_{n-1}	F_n	재방문
151123-afc1..	8	6	6	1.0	.	2	1	1
151123-bie2..	6	4	3	0.7	.	0	0	0

4. 실제 데이터를 이용한 재방문 예측 기계학습 모델

분류기의 시험을 테스트하기 위해, 두 곳의 오프라인 의류 매장 데이터를 수집하였고(표 4), 본 논문에서는 유효 방문 횟수가 많은 매장 A의 데이터를 이용하여 실험을 진행하였다. 충성도가 낮은 고객들의 재방문율을 높이는 것이 중요하다는 연구 목적에 의거하여 방문 횟수가 세 번 이하이고, deny값이 True가 아닌 비단골 고객을 대상으로 샘플링을 진행하였다. 비단골 고객으로의 샘플링 이후 재방문 의사가 있는 방문 패턴 수가 2,689개, 없는 패턴 수가 19,953개가 남았는데 label의 불균형으로 인한 accuracy paradox을 방지하기 위해 재방문 의도에 따른 라벨 비율을 50:50로 조절한 총 5,378개의 내부 이동 패턴을 최종으로 이용하였다.

표 4. 두 오프라인 의류 매장의 데이터 통계

특성	매장 A (스포츠웨어 매장)	매장 B (캐쥬얼 매장)
수집 기간	2015/08/28- 2016/08/31	2015/08/28- 2016/08/31

로그 개수	4,126,399	4,063,446
유효방문 횟수	23,576	16,319
고객 수	18,856	15,464

실험의 완전성을 위해, 서로 다른 1:1 라벨 샘플링을 100번 수행하여 측정한 10-fold cross validation의 평균 정확도를 계산하였고, 특성을 어떻게 택하느냐에 따라 달라지는 예측 결과에 대한 boxplot를 그림 2에서 확인할 수 있다.

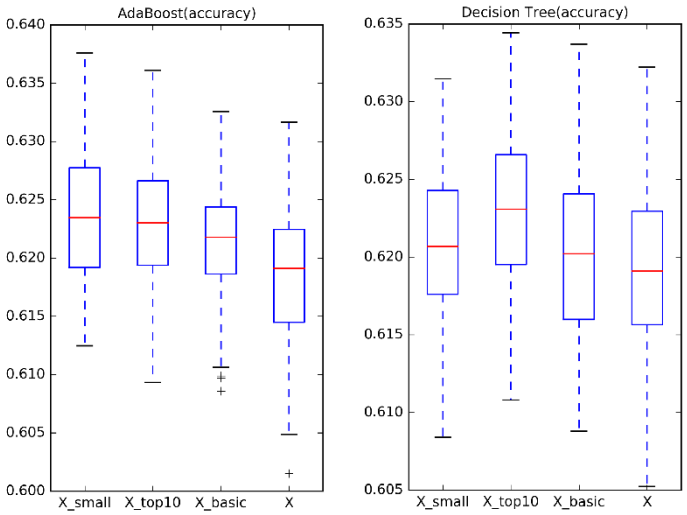

그림 2. 두 가지 classifier를 이용한 재방문율 예측 결과[3] (X_small: 기본 통계 및 요일 특성 14개, X_top10: X에서 중요도 순으로 뽑은 10개 특성, X_basic: 기본 통계 특성 7개, X: 모든 137개 특성을 이용)

본 실험 결과에서는 이동 동선에 관련한 n-gram 특성 이 재방문 예측 정확도 증가를 보장하지 않음을 나타내 주고 있으나, 추후 연구에서 frequent sequence 등을 이용하여 sparsity를 해결하면 더 좋은 결과를 얻을 것으로 생각된다[5].

5. 향후 연구

재방문 예측 분류기 학습을 위해 현재는 표 2에 소개된 동선 스케쥴 데이터 중에서 통계적, 간단한 트라젝토리 패턴에 해당되는 특성을 추출하는 방법을 이용하였다. 마찬가지로, 동선 스케쥴을 Time-series 혹은 sequence data로 압축하여 이용하거나 전처리 후 frequent patterns을 추출하여 이용하는 방법을 사용할 수 있다[6,7]. 하지만, 이런들 역시 원 자료에 포함된 정보들의 일부만을 이용하여 학습하는 방법이다. 따라서 동선 스케쥴과 재방문 라벨을 직접적인 입력으로 받는 지도 학습 모델이나 인공 신경망 모델을 제안하는 것이 앞으로 연구할 과제이다.

6. 결론

고객의 매장 재방문 여부를 예측하기 위해서는 고객의 정보, 구매 이력, 소셜 네트워크에 더불어 매장 내부 이동 패턴 등을 필요로 할 것이다. 하지만 다방면의 데이터를 수집하는 것은 매우 힘들며, 고객의 재방문률이 낮은 오프라인 매장의 특성상 매장에 새로운 고객이 방문할 빈도가 높다. 매장 내에서 움직인 패턴은 제한된 정보이지만, 이를 통해 만들어진 분류기는 고객에 히스토리에 상관없이 이용될 수 있다. 학습된 분류기를 통해 실시간으로 고객의 재방문율을 예측하고 매장에서 퇴장하기 전 할인 쿠폰 지급 등이 가능해진다면 오프라인 매장의 충성 고객 비율이 높아질 것으로 기대된다.

사사

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 보통신기술진흥센터의 지원을 받아 수행된 연구임(No. R0101-16-0054, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)

참고문헌

[1] David Rekuc., "Study: Why 92% of Retail Purchases Still Happen Offline", *Ripen ecommerce*, http://bit.ly/2dxO7zz, 2015.

[2] Yan et al., "Customer Revisit Intention to Restaurants: Evidence from Online Reviews", *Information Systems Frontiers*, 17(3): 645-657, 2015.

[3] Liu et al., "Repeat Buyer Prediction for E-Commerce", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 155-164, 2016.

[4] Sojin Kim., "Analyzing Characteristics of Customers by Using Wi-Fi Log Data", *Master's thesis, KAIST*, 2016

[5] Lee et al., "Mining Discriminative Patterns for Classifying Trajectories on Road Networks", *IEEE Transactions on Knowledge and Data Engineering* 23(5): 713-726, 2011

[6] Esling et al., "Time-series Data Mining", *ACM Computing Surveys*, 45(1) No.12: 2012

[7] Xing et al., "A Brief Survey on Sequence Classification", *ACM SIGKDD Explorations Newsletter*, 12(1): 40-48, 2010.

[3] 3가지 classifier와 5가지 scoring measure를 이용한 전체 실험 결과는 다음 링크(http://bit.ly/2eedLKO)에서 확인할 수 있다.