체크인 시퀀스 기반의 next POI 추천 시스템을 위한 네거티브 샘플링 방법

김예빈 [1], 배홍균 [2], 김상욱 [3*]
[1] 한양대학교 인공지능학과 석사과정
[2] 한양대학교 컴퓨터소프트웨어학과 박사과정
[3] 한양대학교 컴퓨터소프트웨어학과 교수

{kyeb98, hongkyun, wook}@hanyang.ac.kr

A Negative Sampling Method for Next POI Recommender Systems Based on Check-in Sequences

Ye-Been Kim[1], Hong-Kyun Bae[2], Sang-Wook Kim[3]
[1]Dept. of Artificial Intelligence, Hanyang University
[2,3]Dept. of Computer Science, Hanyang University

요 약

최근 위치 기반 장소 (POI) 추천 서비스가 많이 사용되면서, 사용자의 이전 방문지들에 대한 체크인 시퀀스를 기반으로 현재 (다음으로) 방문할 법한 POI 를 찾아 사용자에게 추천하는, next POI 추천 시스템에 관한 연구가 활발히 진행되고 있다. 하지만, 기존 연구들의 경우 next POI 추천을 위한 모델 학습 시, 사용자의 네거티브 POIs 에 관한 정교한 샘플링 없이 사용자 선호도를 추론해왔다. 본 연구에서는, 사전 학습된 별도의 사용자 선호도 추론 모델을 통해 사용자의 네거티브 POI 로서 쉽게 분류되기 어려운 하드 네거티브 POIs 를 찾고, 이들을 위주로 수행되는 하드 네거티브 샘플링 방법을 새롭게 제안한다. 우리는 실 세계 데이터셋을 이용한 실험을 통해, 제안 방안이 기존 연구들에서 사용되어 온 랜덤 네거티브 샘플링 방법 대비 recall@5 기준, 최대 16.4%까지 추천 정확도를 향상시킬 수 있음을 확인하였다.

1. 서론

최근 네이버 스마트 어라운드, 포스퀘어 등 위치 기반 장소 추천 서비스의 사용이 증가하고 있다. 이러한 서비스에서 사용자는 자신이 방문했던 장소들에 대한 체크인 기록을 저장하고, 저장된 체크인 기록 및 현 위치를 바탕으로 사용자가 방문할 법한 장소 (Point-Of-Interest, POI)를 추천받을 수 있다.

POI 추천 시스템은 크게 둘로 나뉘는데, 사용자가 방문했던 장소 셋(set)을 기반으로 유저가 체크인 한 적 없던 POIs 중 유저가 선호할 법한 POIs 를 찾는 POI 추천과 사용자의 체크인 시퀀스(sequence)를 이용해 사용자가 현재 방문지 다음 방문할 법한 POI 를 잘 찾는 POI 추천(이하, next POI 추천)으로 나뉜다. 본 논문은 이 중 체크인 시퀀스를 이용하는 추천 방법에 초점을 둔다. Next POI 추천의 경우 사용자의 최근 방문이 다음 방문에 영향을 미친다는 직관을 바탕으로, 시퀀스 내 사용자의 방문 패턴을 분석하여 현재 방문지 다음으로 방문할 법한 POI 를 추천한다. Next POI 추천을 위한 기존 연구들 [1,2,3,4,5] 에서는 사용자별 체크인 시퀀스 내 각 POI 에 대해서 next POI 를 예측하도록 학습할 때, 실제로 다음으로 방문한 각 POI (포지티브 POI)에 대한 예측점수가 높도록, 방문하지 않은 나머지 POIs (네거티브 POI)에 대한 예측점수는 낮도록 학습한다.

이러한 경우 체크인 시퀀스 내 각 POI 에 대한 모든 네거티브 POIs 를 학습하는데 시간이 많이 걸리는 문제가 있어 네거티브 POIs 중 일부 POI 만 샘플링하여 네거티브 POI 로 학습하는 네거티브 샘플링 방법을 적용하여 해결하였다.

Next POI 추천을 위한 기존 연구들의 경우 사용자가 방문하지 않았던 POI 중 무작위로 샘플링하는 랜덤 샘플링 방법[1, 2, 6]을 사용하거나 포지티브 POI 와의 지리적 거리가 짧은 POI 를 선택하는 방법[3] 등을 이용해 휴리스틱하게 네거티브 샘플을 선택하였다.

* 교신 저자

하지만 기존 연구들의 경우 사용자 및 이전 POI 별로 네거티브로 학습해야 하는 중요한 POI 를 구별하여 샘플링하지 못한다는 문제가 있다. 사용자 및 이전 POI 에 따라 모델이 예측한 네거티브 POI 의 예측 점수가 다른데 네거티브 POI 중 사용자의 선호도가 낮은데도 불구하고 예측점수가 높게 평가된 POI 가 있다면 사용자의 네거티브 POI 로서 쉽게 분류되기 어려운 하드 네거티브 POI 임을 의미하므로, 이러한 샘플들을 위주로 네거티브 샘플링을 수행하여 모델 학습을 수행할 경우 더 정확한 사용자 선호도 추론이 가능해질 것이다.

따라서 본 연구에서는 사전 학습된 추천 모델에서 예측 점수가 높은 POI 일수록 샘플링 확률을 더 높게 주는 방법을 제안한다. 실험에서 제안 방안을 적용해 보았을 때 예측점수가 높을수록 높은 확률로 샘플링 한 방법이 NYC 데이터셋 기준 16.4%, TKY 데이터셋 기준 7.8% 성능향상을 보였다.

2. 관련 연구

[7]은 학습중인 모델의 각 에폭마다 예측점수가 높은 네거티브 아이템을 네거티브 샘플로 선택한다. 이 제안한 방법은 매 에폭마다 예측모델의 예측점수가 높은 아이템을 선택하는 것이고, 우리의 제안 방법은 사전 학습된 모델의 예측점수에 따라 샘플링 확률을 다르게 준다는 점에서 차이가 있다.

Next POI 추천 연구에서 사용된 네거티브 샘플링 방법은 <표 1>과 같다. [1], [2]과 [6]은 랜덤 샘플링을 이용하였고, [3]은 모든 사용자들의 체크인 횟수가 많은 인기있는 아이템은 사용자가 알고 있음에도 불구하고 구매하지 않은 아이템일 확률이 높다는 가정 아래 체크인 횟수가 많을수록 더 샘플링 확률을 높게 하는 [8]의 연구를 next POI 추천에 적용해보았지만 성능이 떨어지는 것을 확인하였다. [3]과 [4]는 현재 위치로부터 멀리 떨어져 있어서 방문하지 않았지만 실제로는 선호할 수 있는 POI 가 네거티브로 학습되는 것을 방지하기 위해 포지티브 POI 와 가까이 있는 POIs 중에서 네거티브 POI 를 선택하였다.

<표 1> 기존 연구에서 사용된 네거티브 샘플링 방법

방법	설명	이용 연구
랜덤 샘플링	사용자가 방문하지 않았던 POIs 중 무작위로 샘플링	STAN[1],CatDM[2], STiSAN [6]
인기도 기반	(모든 사용자들의) 체크인 횟수가 많을 수록 더 많이 샘플링	GeoSAN [3]
거리 기반	Positive poi 와 거리가 가장 가까운 n 개의 POIs 중 무작위로 샘플링	GeoSAN [3]
	Positive POI 와 같은 도시에 있는 POIs 중 샘플링	ATST-LSTM [4]
카테고리 기반	Positive POI 와 다른 카테고리내 POIs 중 샘플링	HCT [5]

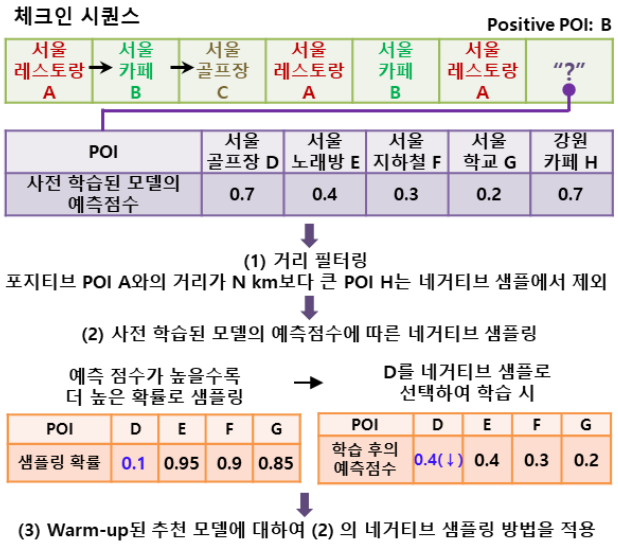
3. 제안 방안

우리의 제안 방안은 (그림 1)과 같이 (1) 포지티브 POI 와 거리가 먼 네거티브 POI 를 네거티브 샘플후보에서 제외한 후 (2) 사전 학습된 모델의 예측점수가 높을수록 샘플링 확률이 높도록 계산하고, (3) 추천 모델을 일정 에폭까지 학습 시킨 뒤 (2)에서 계산한 샘플링 확률에 따라 네거티브 샘플을 선택해 학습한다.

먼저 (1)에서 [3]과 [4]의 연구를 따라, 거리가 일정 거리 이상 떨어진 POI 를 네거티브 샘플 후보에서 제외한다.

(2) 에서는 먼저 학습하고자 하는 추천모델과 동일한 별도의 모델을 랜덤샘플링을 이용하여 사전 학습시킨다. 사전 학습이 완료되면 해당 모델에서 사용자 및 이전 POI 에 따라 예측한 네거티브 POI 에 대한 예측점수를 산출하여 예측점수를 사분위수에 따라 네 개의 그룹으로 나눈다. 예측 점수가 높을수록 더 높은 확률로 샘플링 되도록 예측점수가 1 분위수 안인 네거티브 POI 보다 예측점수가 2 분위수, 3 분위수, 4 분위수 안에 드는 POIs 의 샘플링 확률이 각각 5, 10, 15 % 더 높도록 설정한다.

(3) 에서는 일정 K 에폭까지 랜덤샘플링을 통해 추천모델을 학습시켜 warm-up 된 추천 모델에 대하여 이후 에폭부터 (2)에서 설정한 샘플링 확률에 따라 네거티브 샘플을 선택하여 추천 모델을 학습시킨다.



(그림 1) 제안 방안.

4. 실험

4.1 데이터셋

<표 2> 데이터 셋

	사용자수	POI 수	체크인수	희소성
NYC	1,000	5,136	138,361	97.3%
TKY	500	7,872	100,128	97.4%

실험에서 실세계 데이터 셋인 포스퀘어 뉴욕, 도쿄 데이터셋을 이용하였다. 이용한 데이터셋의 사용자,

<표 3> NYC 데이터셋 결과

네거티브 샘플링 방법	Warm-up 수행 여부	R@5	R@10	R@15	R@20	M@5	M@10	M@15	M@20	G@5	G@10	G@15	G@20
랜덤 샘플링	-	0.3155	0.396	0.44	0.4885	0.1981	0.2088	0.2124	0.2151	0.2274	0.2534	0.2652	0.2767
하드 네거티브 샘플링	X	0.277	0.359	0.3995	0.4265	0.1761	0.1871	0.1903	0.1918	0.2012	0.2278	0.2385	0.2449
하드 네거티브 샘플링	O	**0.3675**	**0.4485**	**0.5065**	**0.5365**	**0.2389**	**0.2495**	**0.2541**	**0.2558**	**0.2712**	**0.2971**	**0.3125**	**0.3196**
이지 네거티브 샘플링	X	0.3005	0.368	0.403	0.434	0.1817	0.1911	0.1937	0.1955	0.2112	0.2334	0.2425	0.2498
이지 네거티브 샘플링	O	0.342	0.427	0.4705	0.507	0.2223	0.2341	0.2375	0.2396	0.2522	0.2801	0.2916	0.3003

<표 4> TKY 데이터셋 결과

네거티브 샘플링 방법	Warm-up 수행 여부	R@5	R@10	R@15	R@20	M@5	M@10	M@15	M@20	G@5	G@10	G@15	G@20
랜덤 샘플링	-	0.1901	0.2674	0.3180	0.3495	0.1148	0.1251	0.1291	0.1308	0.1335	0.1585	0.1719	0.1793
하드 네거티브 샘플링	X	0.1683	0.2253	0.2651	0.2958	0.1009	0.1085	0.1116	0.1133	0.1176	0.1360	0.1465	0.1537
하드 네거티브 샘플링	O	**0.2049**	**0.2844**	**0.3253**	**0.3553**	**0.1163**	**0.1270**	**0.1302**	**0.1319**	**0.1383**	**0.1640**	**0.1748**	**0.1820**
이지 네거티브 샘플링	X	0.1468	0.1976	0.2311	0.2555	0.0928	0.0996	0.1022	0.1036	0.1061	0.1225	0.1314	0.1372
이지 네거티브 샘플링	O	0.1911	0.2664	0.3129	0.3476	0.1109	0.1209	0.1246	0.1265	0.1308	0.1551	0.1674	0.1756

POI 수, 체크인 수는 <표 2>와 같다. 우리는 데이터셋을 시간순으로 정렬한 후 사용자마다 처음 70%를 훈련 데이터셋, 10%를 검증 데이터셋, 가장 최근에 방문한 20%를 테스트 데이터셋으로 사용하였다.

4.2 실험 설정

우리는 추천 모델로 STAN [1] 을 사용하였다. 모델 파라미터들은 STAN 논문의 설정을 적용하였고, 거리 필터링 파라미터 N 을 10km 로 설정하였다. NYC 데이터셋에 대해서는 50 에폭, TKY 데이터셋에 대해 60 에폭까지 학습을 진행하였다. 그리고, 추천 모델의 사전학습 에폭 K 를 NYC 데이터셋에서는 30 에폭, TKY 데이터셋에서는 50 에폭 적용하였다.

4.3 실험 결과

<표 3>은 NYC 데이터셋에 대하여, <표 4>는 TKY 데이터셋에 대하여 본 논문에서 제안하는 샘플링 방법의 실험 효과를 보인다. 먼저, 하드 네거티브 샘플 위주로 샘플링하는 방법 적용시, 사전학습 모델을 이용하지 않으면 랜덤샘플링 방법 대비 NYC 데이터셋 recall@5 기준 -12.2%, TKY 데이터셋 recall@5 기준 -11.5%의 성능 하락을 보였고, 사전학습 모델 이용 시 NYC 데이터셋 기준 16.4%, TKY 데이터셋 기준 7.8% 성능향상을 보였다.

또한 우리는 거리 필터링은 동일하게 적용하고, 예측점수가 낮아 네거티브 POI 로서의 분류가 쉬운 이지 네거티브 샘플 위주로 샘플링을 수행해 보았을 때, 사전학습 모델 이용 시 하드네거티브 샘플링 방법이 이지 네거티브 샘플링 방법 대비 NYC 데이터셋 기준 7.2%, TKY 데이터셋 기준 7.2% 성능이 높음을 확인하였다.

5 결론

기존 next POI 추천연구들의 경우 사용자의 네거티브 POIs 에 관한 정교한 샘플링 없이 사용자 선호도를 추론해왔다. 그래서 본 연구에서 사전 학습된 별도의 사용자 선호도 추론 모델을 통해 사용자의 네거티브 POI 로서 쉽게 분류되기 어려운 하드 네거티브 POIs 를 찾고, 해당 네거티브 POIs 를 위주로 샘플링하는 방법을 새롭게 제안하였다. 실험을 통해, 제안 방안이 기존 연구들에서 주로 사용되어 온 랜덤 네거티브 샘플링 방법 대비 추천 정확도를 향상시킬 수 있음을 확인하였다

감사의 글

본 논문은 (1) 2023 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원과 (No.RS-2022-00155586, 실세계의 다양한 다운스트림 태스크를 위한 고성능 빅 하이퍼그래프 마이닝 플랫폼 개발(SW 스타랩)), (2) 문화체육관광부 및 한국콘텐츠진흥원의 문화기술 연구개발 사업과 (과제명 : 지능형 개인맞춤 재활운동 서비스 기술개발, 과제번호 : SR202104001, 기여율: 00%), (3) 2023 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2020-0-01373, 인공지능대학원지원(한양대학교))

참고문헌

[1] LUO, Yingtao; LIU, Qiang; LIU, Zhaocheng. Stan: Spatio-temporal attention network for next location recommendation. *WWW*. Ljubljana, Slovenia. 2021. p. 2177-2185.

[2] YU, Fuqiang, et al. A category-aware deep model for successive POI recommendation on sparse check-in data. *WWW*. Taipei, Taiwan. 2020. p. 1264-1274.

[3] LIAN, Defu, et al. Geography-aware sequential location recommendation. *SIGKDD*. CA, USA. 2020. p. 2009-2019.

[4] HUANG, Liwei, et al. An attention-based spatiotemporal lstm network for next poi recommendation. *IEEE Transactions on Services Computing*, VOL. 14, NO. 6: p.1585-1597.

[5] ZHANG, Lu, et al. Modeling hierarchical category transition for next POI recommendation with uncertain check-ins. *Information Sciences*, 515: p.169-190.

[6] WANG, En, et al. Spatial-Temporal Interval Aware Sequential POI Recommendation. *ICDE*. Kuala Lumpur, Malaysia. 2022. p. 2086-2098.

[7] ZHANG, Weinan, et al. Optimizing top-n collaborative filtering via dynamic negative item sampling. *SIGIR*. Dublin, Ireland. 2013. p. 785-788.

[8] HIDASI, Balázs, et al. Session-based recommendations with recurrent neural networks. *ICLR*. San Diego, USA. 2016. 1-10.