Challenges on Evaluating Venue Recommendation Approaches

(Position paper)

Pablo Sánchez, **Alejandro Bellogín**
Universidad Autónoma de Madrid
Spain

RecTour @ RecSys, October 2018

Position paper

- Discuss two issues with the community:
  - Evaluation methodologies in traditional venue recommendation
  - How to integrate sequences when evaluating venue recommenders

Position paper

- Discuss two issues with the community:
  - Evaluation methodologies in traditional venue recommendation
  - How to integrate sequences when evaluating venue recommenders

- However, thanks to the reviewers, a third issue arised:
  - Check-ins vs tourists: is there a realistic tourism dataset?
  - Usefulness of LBSN datasets? Foursquare, Gowalla, etc.
  - ...
  - Left as future work

Challenge 1: evaluation methodology

- Two possibilities when building the test set
  - Only new venues
  - No filter: known (previously visited) venues by the user

UAM
Universidad Autónoma
de Madrid

Escuela
Politécnica
Superior

Challenge 1: evaluation methodology

- Two possibilities when building the test set
  - Only new venues
  - No filter: known (previously visited) venues by the user

test split

Challenge 1: evaluation methodology

- Two possibilities when building the test set
  - Only new venues
  - No filter: known (previously visited) venues by the user
- Each possibility translates into different recommendation tasks
  - Recommending new places [Bothorel et al 2018]
  - Recommending what a user will visit next (without considering novelty)
- Important
  - For reproducibility purposes
  - Choice of experimental conditions

Experiments on challenge 1

Recommender	Test with new venues P	R	NDCG	MAP	Test with known venues P	R	NDCG	MAP
Rnd	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Pop	0.039	0.076	0.063	0.030	0.054	0.082	0.079	0.036
Training	0.000	0.000	0.000	0.000	**†0.120**	**†0.190**	0.186	0.100
AvgDis	0.001	0.001	0.001	0.001	0.003	0.006	0.007	0.005
AvgDisFreq	0.001	0.002	0.001	0.001	0.003	0.007	0.008	0.005
PGN	0.041	0.082	0.073	0.036	0.070	0.112	0.124	0.065
UB	0.045	0.088	0.078	0.039	0.110	0.167	0.178	0.098
IB	0.036	0.069	0.063	0.032	0.108	0.156	0.175	0.098
HKV	0.043	0.087	0.076	0.039	0.105	0.158	0.170	0.093
IRenMF	0.044	0.089	0.077	0.039	0.100	0.151	0.164	0.090
IRenMFFreq	**†0.047**	**†0.094**	**†0.082**	**†0.042**	**0.117**	**0.181**	**†0.194**	**†0.109**

- Popularity bias in ‘new venues’ scenario
- *Training* bias (this baseline is hard to beat) in ‘known venues’ scenario

Challenge 1: discussion

- How the test split is created is critical
- If known venues are included, a baseline similar to the one used here (Training) should be added in the comparison
- Helps reproducibility and fair reasoning about the results

Challenge 2: evaluating with sequences

- Sequences prevalent in recommendation nowadays [Quadrana et al 2018]
  - Tourism as a special case: a route is a sequence of venues
- Can we also consider the order (in test) when evaluating?

Challenge 2: evaluating with sequences

- Sequences prevalent in recommendation nowadays [Quadrana et al 2018]
  - Tourism as a special case: a route is a sequence of venues
- Can we also consider the order (in test) when evaluating?
  - Proposal: use LCS (Longest Common Subsequence) algorithm

$$L[i, j] = \begin{cases} 0 & \text{if i = 0 or j = 0} \\ L[i-1, j-1] + 1 & \text{if } i, j > 0 \text{ and } x_i = y_j \\ \max(L[i, j-1], L[i-1, j]) & \text{if } i, j > 0 \text{ and } x_i \neq y_j \end{cases}$$

	∅	A	G	G	T
∅	0	0	0	0	0
G	0	0	1	1	1
C	0	0	1	1	1
G	0	0	1	2	2
T	0	0	1	2	3

Challenge 2: evaluating with sequences

- Sequences prevalent in recommendation nowadays [Quadrana et al 2018]
  - Tourism as a special case: a route is a sequence of venues
- Can we also consider the order (in test) when evaluating?
  - Proposal: use LCS (Longest Common Subsequence) algorithm
  - Three variations (at cutoff *N*, where R_u is the recommended ranking)

$$LCSP(R_u, T_u) = \frac{\text{lcs}(R_u, T_u)}{N}$$

Based on precision

$$LCSR(R_u, T_u) = \frac{\text{lcs}(R_u, T_u)}{|T_u|}$$

Based on recall

$$LCS(R_u, T_u) = \frac{\text{lcs}(R_u, T_u)^2}{N \cdot |R_u|}$$

Normalized LCS

Challenge 2: evaluating with sequences

- Sequences prevalent in recommendation nowadays [Quadrana et al 2018]
  - Tourism as a special case: a route is a sequence of venues
- Can we also consider the order (in test) when evaluating?
  - Proposal: use LCS (Longest Common Subsequence) algorithm
  - Three variations (at cutoff N, where R_u is the recommended ranking)
  - We can capture how similar the recommendation is to the test with respect to the order followed by the user

Experiments on challenge 2

- Included two skylines that use the test

Recommender	Test with new venues NDCG	LCS	LCSP	LCSR	Test with known venues NDCG	LCS	LCSP	LCSR
Rnd	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Pop	0.063	0.008	0.034	0.071	0.079	0.009	0.046	0.075
Training	0.000	0.000	0.000	0.000	0.186	†**0.034**	0.090	†**0.157**
AvgDis	0.001	0.000	0.001	0.001	0.007	0.001	0.002	0.006
AvgDisFreq	0.001	0.000	0.001	0.001	0.008	0.001	0.003	0.006
PGN	0.073	0.009	0.037	0.077	0.124	0.013	0.059	0.101
UB	0.078	0.009	0.039	0.081	0.178	0.021	0.086	0.142
IB	0.063	0.008	0.032	0.064	0.175	0.019	0.082	0.130
HKV	0.076	0.009	0.038	0.080	0.170	0.019	0.082	0.135
IRenMF	0.077	0.010	0.039	0.083	0.164	0.018	0.079	0.130
IRenMFFreq	†**0.082**	†**0.010**	†**0.041**	†**0.087**	†**0.194**	**0.023**	†**0.092**	**0.154**
TestInvOrder	0.978	0.225	0.100	0.356	0.985	0.162	0.100	0.287
TestOrder	0.978	0.932	0.468	0.932	0.985	0.910	0.569	0.910

- The LCS-based metrics are successfully capturing the test order
- The Training baseline and IRenMF perform very well in terms of order

Challenge 2: discussion

- Our proposal seems to be able to capture how well a recommender matches the order followed by the user

- The results evidence that there is still room for improvement
  - In the future, we want to test these metrics with algorithms that recommend sequences of items

- Is it possible to generalize this definition to more complex metrics, such as NDCG?

Conclusions

- Known vs new venues in test set
  - Different tasks with different starting hypotheses
  - Results change dramatically in each situation
  - Hence, the experimental design should be properly described
- Capturing user sequences in evaluation
  - Metrics based on LCS successfully assess the similarity between the recommended list and the user test
- What happens with check-in datasets? Are they useful to investigate tourism recommenders?

- Source code: https://bitbucket.org/PabloSanchezP/TempCDSeqEval

Thank you

Challenges on Evaluating Venue Recommendation Approaches

(Position paper)

Pablo Sánchez, Alejandro Bellogín

Universidad Autónoma de Madrid

Spain

RecTour @ RecSys, October 2018

Universidad Autónoma de Madrid

Escuela Politécnica Superior

References

- [Bothorel et al 2018] Location recommendation with Social Media Data. 2018. *Social Information Access, 624-653*.
- [Quadrana et al 2018] Sequence-Aware Recommender Systems. 2018. *ACM Comput. Surv.* 51, 4, 66:1-66:36.
- [Yang et al 2015] NationTelescope: Monitoring and visualizing large-scale collective behavior in LBSNs. 2015. *J. Network and Computer Applications* 55, 170–180.

Dataset

Table 1: Description of the temporal partition evaluated created based on the Foursquare dataset, where U, I, and C denote the number of users, items, and check-ins.

Check-in period	U	I	C	Density	C/U	C/I
Apr'12-Sep'13	267k	3.6M	33M	0.0034%	123.596	9.16
Training: May-Oct '12	202k	1.1M	4.7M	0.0021%	23.267	4.278
Test: Nov '12	150k	352k	831k	0.0017%	5.540	2.361

https://sites.google.com/site/yangdingqi/home/foursquare-dataset

Full experiments: Istanbul

Recommender	Test with new venues							Test with known venues						
	P	R	NDCG	MAP	LCS	LCSP	LCSR	P	R	NDCG	MAP	LCS	LCSP	LCSR
Rnd	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Pop	0.039	0.076	0.063	0.030	0.008	0.034	0.071	0.054	0.082	0.079	0.036	0.009	0.046	0.075
Training	0.000	0.000	0.000	0.000	0.000	0.000	0.000	†0.120	†0.190	0.186	0.100	†0.034	0.090	†0.157
AvgDis	0.001	0.001	0.001	0.001	0.000	0.001	0.001	0.003	0.006	0.007	0.005	0.001	0.002	0.006
AvgDisFreq	0.001	0.002	0.001	0.001	0.000	0.001	0.001	0.003	0.007	0.008	0.005	0.001	0.003	0.006
PGN	0.041	0.082	0.073	0.036	0.009	0.037	0.077	0.070	0.112	0.124	0.065	0.013	0.059	0.101
UB	0.045	0.088	0.078	0.039	0.009	0.039	0.081	0.110	0.167	0.178	0.098	0.021	0.086	0.142
IB	0.036	0.069	0.063	0.032	0.008	0.032	0.064	0.108	0.156	0.175	0.098	0.019	0.082	0.130
HKV	0.043	0.087	0.076	0.039	0.009	0.038	0.080	0.105	0.158	0.170	0.093	0.019	0.082	0.135
IRenMF	0.044	0.089	0.077	0.039	0.010	0.039	0.083	0.100	0.151	0.164	0.090	0.018	0.079	0.130
IRenMFFreq	†0.047	†0.094	†0.082	†0.042	†0.010	†0.041	†0.087	0.117	0.181	†0.194	†0.109	0.023	†0.092	0.154
TestInvOrder	0.468	0.932	0.978	0.967	0.225	0.100	0.356	0.569	0.910	0.985	0.978	0.162	0.100	0.287
TestOrder	0.468	0.932	0.978	0.967	0.932	0.468	0.932	0.569	0.910	0.985	0.978	0.910	0.569	0.910

UAM Universidad Autónoma de Madrid

Escuela Politécnica Superior

Full experiments: Jakarta

Recommender	Test with new venues P	R	NDCG	MAP	LCS	LCSP	LCSR	Test with known venues P	R	NDCG	MAP	LCS	LCSP	LCSR
Rnd	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Pop	0.029	0.076	0.070	0.044	0.008	0.026	0.073	0.044	0.087	0.091	0.056	0.009	0.038	0.082
Training	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.102	0.196	0.171	0.096	†0.034	0.078	0.165
AvgDis	0.001	0.002	0.002	0.001	0.000	0.001	0.002	0.003	0.008	0.007	0.005	0.001	0.002	0.007
AvgDisFreq	0.001	0.002	0.001	0.001	0.000	0.001	0.002	0.004	0.010	0.009	0.006	0.001	0.003	0.009
PGN	0.030	0.078	0.072	†0.045	0.008	0.027	0.075	0.056	0.108	0.114	0.069	0.012	0.047	0.100
UB	0.036	0.085	0.075	0.043	0.009	0.032	0.081	0.081	0.141	0.146	0.083	0.019	0.065	0.124
IB	0.019	0.045	0.038	0.021	0.005	0.017	0.043	†0.120	†0.212	†0.222	†0.141	0.026	†0.088	†0.172
HKV	0.035	0.084	0.071	0.039	0.009	0.032	0.080	0.078	0.137	0.138	0.078	0.016	0.063	0.121
IRenMF	0.033	0.081	0.071	0.041	0.009	0.030	0.078	0.076	0.135	0.136	0.078	0.016	0.062	0.121
IRenMFFreq	†0.036	†0.092	†0.077	0.044	†0.010	†0.033	†0.088	0.110	0.199	0.193	0.115	0.024	0.084	0.170
TestInvOrder	0.387	0.923	0.963	0.947	0.299	0.100	0.427	0.492	0.912	0.977	0.966	0.223	0.100	0.348
TestOrder	0.387	0.923	0.963	0.947	0.923	0.387	0.923	0.492	0.912	0.977	0.966	0.912	0.492	0.912