A Methodology for the Offline Evaluation of Recommender Systems in a User Interface with Multiple Carousels

NICOLÒ FELICIONI, Politecnico di Milano, Italy
MAURIZIO FERRARI DACREMA, Politecnico di Milano, Italy
PAOLO CREMONESI, Politecnico di Milano, Italy

CCS Concepts: • **Information systems** → **Collaborative filtering**; **Recommender systems**; • **General and reference** → **Evaluation**.

Additional Key Words and Phrases: Recommender Systems; User Interface; Evaluation

ACM Reference Format:
Nicolò Felicioni, Maurizio Ferrari Dacrema, and Paolo Cremonesi. 2021. A Methodology for the Offline Evaluation of Recommender Systems in a User Interface with Multiple Carousels. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21 Adjunct), June 21–25, 2021, Utrecht, Netherlands.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3450614.3461680

This is the additional material for the paper "A Methodology for the Offline Evaluation of Recommender Systems in a User Interface with Multiple Carousels".

It contains the results of the experiments done on MovieLens10M with TopPop (Table 1) and SLIM ELasticNet (Table 2) as the fixed carousel, as well as the experiments on the Netflix dataset (TopPop in Table 3, SLIM ElasticNet in Table 4).

Manuscript submitted to ACM

1 MOVIELENS 10M

	Individual			Carousel				Improvement on TopPop		MAP rank		
	PREC	MAP	NDCG	PREC	MAP	NDCG	NDCG 2D	Individual	Carousel	Individual	Carousel	Δ rank
TopPop	0.0975	0.0709	0.0983	–	–	–	–	–	–	–	–	–
UserKNN CF	0.2343	0.2251	0.2815	0.1338	0.0889	0.1911	0.2492	+217.53%	+68.59%	2	5	-3
ItemKNN CF	0.1885	0.1728	0.2122	0.1164	0.0812	0.1662	0.2093	+143.85%	+54.00%	9	10	-1
$P^3\alpha$	0.1646	0.1414	0.1915	0.0897	0.0687	0.1451	0.1710	+99.47%	+30.29%	13	13	0
$RP^3\beta$	0.1886	0.1686	0.2160	0.1104	0.0779	0.1632	0.2035	+137.81%	+47.83%	10	11	-1
$EASE^R$	0.2260	0.2070	0.2566	0.1281	0.0855	0.1799	0.2293	+192.10%	+62.10%	5	8	-3
SLIM BPR	0.2274	0.2159	0.2699	0.1291	0.0872	0.1868	0.2414	+204.64%	+65.40%	3	7	-4
SLIM ElasticNet	0.2460	0.2340	0.2856	0.1399	0.0915	0.1931	0.2530	+230.21%	+73.48%	1	2	-1
MF BPR	0.1759	0.1502	0.1882	0.1088	0.0773	0.1581	0.1932	+111.85%	+46.66%	12	12	0
MF FunkSVD	0.2039	0.1748	0.2307	0.1327	0.0884	0.1859	0.2373	+146.64%	+67.71%	8	6	+2
PureSVD	0.2217	0.2060	0.2527	0.1329	0.0890	0.1824	0.2357	+190.59%	+68.74%	6	4	+2
NMF	0.1872	0.1613	0.1974	0.1319	0.0895	0.1792	0.2281	+127.51%	+69.76%	11	3	+8
IALS	0.2329	0.2152	0.2539	0.1425	0.0942	0.1895	0.2446	+203.64%	+78.75%	4	1	+3
ItemKNN CBF	0.0113	0.0052	0.0079	0.0541	0.0544	0.1022	0.1007	-92.63%	+3.15%	14	14	0
ItemKNN CFCBF	0.1950	0.1787	0.2151	0.1198	0.0827	0.1696	0.2148	+152.06%	+56.77%	7	9	-2

Table 1. Comparison of the MovieLens 10M accuracy metrics with individual and carousel evaluation at recommendation list length of 10. Note that in the carousel evaluation there will be two recommendation lists. The improvement over the TopPop carousel is computed with MAP.

	Individual			Carousel				Improvement on SLIM EN		MAP rank		
	PREC	MAP	NDCG	PREC	MAP	NDCG	NDCG 2D	Individual	Carousel	Individual	Carousel	Δ rank
SLIM EN	0.2460	0.2340	0.2856	–	–	–	–	–	–	–	–	–
TopPop	0.0975	0.0709	0.0983	0.1399	0.1895	0.2967	0.2939	-69.7%	+4.8%	13	13	0
UserKNN CF	0.2343	0.2251	0.2815	0.1528	0.1955	0.3137	0.3225	-3.8%	+8.1%	1	3	-2
ItemKNN CF	0.1885	0.1728	0.2122	0.1455	0.1921	0.3015	0.3034	-26.2%	+6.3%	8	9	-1
$P^3\alpha$	0.1646	0.1414	0.1915	0.1433	0.1912	0.3009	0.3021	-39.6%	+5.7%	12	10	+2
$RP^3\beta$	0.1886	0.1686	0.2160	0.1430	0.1908	0.3014	0.3026	-28.0%	+5.5%	9	11	-2
$EASE^R$	0.2260	0.2070	0.2566	0.1430	0.1899	0.3017	0.3012	-11.5%	+5.1%	4	12	-8
SLIM BPR	0.2274	0.2159	0.2699	0.1490	0.1937	0.3084	0.3138	-7.7%	+7.2%	2	6	-4
MF BPR	0.1759	0.1502	0.1882	0.1479	0.1937	0.3040	0.3069	-35.8%	+7.2%	11	5	+6
MF FunkSVD	0.2039	0.1748	0.2307	0.1560	0.1979	0.3148	0.3248	-25.3%	+9.5%	7	2	+5
PureSVD	0.2217	0.2060	0.2527	0.1471	0.1924	0.3039	0.3061	-12.0%	+6.4%	5	7	-2
NMF	0.1872	0.1613	0.1974	0.1484	0.1938	0.3037	0.3064	-31.1%	+7.2%	10	4	+6
IALS	0.2329	0.2152	0.2539	0.1592	0.1998	0.3101	0.3174	-8.1%	+10.5%	3	1	+2
ItemKNN CBF	0.0113	0.0052	0.0079	0.1264	0.1826	0.2875	0.2765	-97.8%	+1.0%	14	14	0
ItemKNN CFCBF	0.1952	0.1790	0.2174	0.1460	0.1923	0.3021	0.3044	-23.5%	+6.4%	6	8	-2

Table 2. Comparison of the MovieLens10M accuracy metrics with individual and carousel evaluation (with SLIM EN fixed as the first carousel) at recommendation list length of 10. Note that in the carousel evaluation there will be two recommendation lists. The improvement over the SLIM EN carousel is computed with MAP.

2 NETFLIX

	Individual			Carousel				Improvement on TopPop		MAP rank		
	PREC	MAP	NDCG	PREC	MAP	NDCG	NDCG 2D	Individual	Carousel	Individual	Carousel	Δ rank
TopPop	0.0984	0.0617	0.0606	–	–	–	–	–	–	–	–	–
UserKNN CF	0.2301	0.1906	0.1887	0.1369	0.0710	0.1305	0.1799	+209.12%	+76.25%	3	7	-4
ItemKNN CF	0.2026	0.1579	0.1548	0.1246	0.0656	0.1165	0.1565	+156.14%	+62.90%	6	8	-2
$P^3\alpha$	0.1708	0.1244	0.1391	0.1010	0.0556	0.1048	0.1360	+101.74%	+38.11%	12	12	0
$RP^3\beta$	0.1878	0.1410	0.1613	0.1179	0.0628	0.1192	0.1607	+128.76%	+55.84%	10	9	+1
$EASE^R$	0.2444	0.2014	0.2025	0.1448	0.0747	0.1353	0.1886	+226.64%	+85.34%	2	4	-2
SLIM BPR	0.2019	0.1512	0.1626	0.1175	0.0622	0.1177	0.1540	+145.27%	+54.50%	8	10	-2
SLIM ElasticNet	0.2556	0.2115	0.2205	0.1545	0.0790	0.1468	0.2077	+242.99%	+96.12%	1	1	0
MF BPR	0.1753	0.1271	0.1260	0.1139	0.0618	0.1062	0.1375	+106.22%	+53.39%	11	11	0
MF FunkSVD	0.1991	0.1481	0.1615	0.1372	0.0724	0.1297	0.1769	+140.19%	+79.82%	9	6	+3
PureSVD	0.2343	0.1893	0.1920	0.1469	0.0765	0.1347	0.1876	+207.12%	+89.89%	4	3	+1
NMF	0.2131	0.1658	0.1695	0.1377	0.0724	0.1265	0.1736	+168.87%	+79.83%	5	5	0
IALS	0.2062	0.1562	0.1512	0.1500	0.0787	0.1312	0.1785	+153.34%	+95.40%	7	2	+5

Table 3. Comparison of the Netflix accuracy metrics with individual and carousel evaluation (with a TopPop fixed as the first carousel) at recommendation list length of 10. Note that in the carousel evaluation there will be two recommendation lists. The improvement over the TopPop carousel is computed with MAP.

	Individual			Carousel				Improvement on SLIM EN		MAP rank		
	PREC	MAP	NDCG	PREC	MAP	NDCG	NDCG 2D	Individual	Carousel	Individual	Carousel	Δ rank
SLIM EN	0.2556	0.2115	0.2205	–	–	–	–	–	–	–	–	–
TopPop	0.0984	0.0617	0.0606	0.1545	0.1511	0.2326	0.2420	-70.84%	+9.93%	12	10	+2
UserKNN CF	0.2301	0.1906	0.1887	0.1665	0.1557	0.2453	0.2653	-9.88%	+13.29%	2	5	-3
ItemKNN CF	0.2026	0.1579	0.1548	0.1636	0.1547	0.2408	0.2578	-25.32%	+12.53%	5	6	-1
$P^3\alpha$	0.1708	0.1244	0.1391	0.1566	0.1516	0.2358	0.2486	-41.18%	+10.31%	11	9	+2
$RP^3\beta$	0.1878	0.1410	0.1613	0.1540	0.1502	0.2351	0.2476	-33.30%	+9.27%	9	12	-3
$EASE^R$	0.2444	0.2014	0.2025	0.1587	0.1509	0.2393	0.2535	-4.77%	+9.81%	1	11	-10
SLIM BPR	0.2019	0.1512	0.1626	0.1652	0.1558	0.2416	0.2587	-28.49%	+13.38%	7	4	+3
MF BPR	0.1753	0.1271	0.1260	0.1670	0.1572	0.2418	0.2589	-39.88%	+14.41%	10	2	+8
MF FunkSVD	0.1991	0.1481	0.1615	0.1671	0.1568	0.2462	0.2667	-29.97%	+14.07%	8	3	+5
PureSVD	0.2343	0.1893	0.1920	0.1637	0.1542	0.2417	0.2586	-10.46%	+12.21%	3	7	-4
NMF	0.2131	0.1658	0.1695	0.1596	0.1524	0.2388	0.2536	-21.61%	+10.88%	4	8	-4
IALS	0.2062	0.1562	0.1512	0.1773	0.1632	0.2460	0.2667	-26.14%	+18.77%	6	1	+5

Table 4. Comparison of the Netflix accuracy metrics with individual and carousel evaluation (with SLIM ElasticNet fixed as the first carousel) at recommendation list length of 10. Note that in the carousel evaluation there will be two recommendation lists. The improvement over the SLIM ElasticNet carousel is computed with MAP.