文章编号:1007-1423(2018)24-0011-06 DOI:10.3969/j.issn.1007-1423.2018.24.003

推荐系统评估研究综述

刘春霞,武玲梅,谢小红

(广西师范学院计算机与信息工程学院,南宁 530299)

摘要:

推荐系统逐渐流行于各大商业领域与科研领域,目前已有大量的相关研究。然而大部分学者仅专注于推荐算法的研究而忽视推荐系统评估的研究。评估是检验推荐系统性能的手段,对于推荐系统的发展具有重要的引导意义。着眼于评估的研究,对著名杂志 ACM TOIS 近五年发表的相关的文献进行调查、分析、总结,指出评估研究的现状,阐述三种实验的优势与局限性,并着重对现有的评测指标进行系统的回顾。根据目前评估研究存在的问题,提出五个未来可研究的方向。

关键词:

信息过载;推荐系统;系统评估;评测指标

基金项目:

国家自然科学基金资助项目(No.61672177)

0 引言

这是一个“互联网+”的时代,时代背景造就了信息过载(Information Overload)[1]的现状。推荐系统(Recommender System,RS)通过信息过滤,能够给用户提供所需的信息,满足用户的个性化需求,逐渐在各个应用领域崭露头角。这促使了学者不断进行研究,推荐算法层出叠现,而针对 RS 评估的研究却是屈指可数。学者在研究时如何清晰地鉴别算法的优劣,如何在众多评价指标中选用适当的指标对推荐算法的优劣进行评估,国内外学者目前还未能达成统一定论。本文在分析 RS 评估研究现状的基础上,对评估指标进行系统回顾,并提供了可研究的几个方向。

1 RS评估的研究现状

Anderson 在《长尾理论》中预言,“我们正迈入推荐时代。”学者们对 RS 的研究产生的浓厚的兴趣,而针对其评估的研究却寥寥无几。为客观分析其研究现状,本文选取了著名期刊 *ACM Transaction On Information Systems(ACM TOIS)*近五年间与推荐相关的论文进行调查分析。表 1 列出了其中的 10 篇文献。调查发现,在选取指标对推荐算法进行评估时,部分学者仍对 R 评估指标认识并不全面,在验证算法时仅选取很少的指标,对其他指标的表现缄口不言。另外,部分学者都是主观地选择评估指标去评估算法,在所属文献中,对实验结果解释不足。

表 1 ACM TOIS 近 5 年推荐相关文献

参考文献	测量指标(简称)
LIANG 等人[2]	MAE,RMSE,Recall,precision,MAP
Kalamatianos G 等人[3]	NDCG, precision
CHENYI ZHANG 等人[4]	RMSE,MAE
DA CAO 等人[5]	MAE,RMSE,Recall,DCG,NDCG
HAO M 等人[6]	MAE RMSE
ZHIYONG CHENG 等人[7]	Precision,AP,NDCG
YONG GE 等人[8]	Precision, AP, RMSE
LOC DO 等人[9]	RMSE
RANA FORSATI 等人[10]	MAE, RMSE, AP, NDCG
LEI SHI 等人[11]	Precision, MAP

2 RS实验方法

不管是从哲学观点还是统计学角度,人们对于实验研究本身一直持有非常审慎的态度[12]。一个 RS 最终上线,可以通过线下评估、用户调研评估、线上评估

三种方式对 RS 进行评估。线下评估一般是通过数据集模拟用户与 RS 的交互行为，是科研的首选。用户调查需要一组测试对象，他们须根据需求与系统进行交互。在他们完成交互任务时，我们须要观察并记录他们的行为，根据测试的各个阶段提出定性的问题，用于收集不能直观获取的数据。最后，通过分析收集的数据了解系统的性能。线上评估实际上是通过设计几种方案让线上用户进行体验，通过用户的反馈来判断优劣。AB(All-Between)测试作为一种验证性工具，是常用的线上评估方法。三种实验方法的优缺点如表 2 所示。

表 2 三种实验方法优缺点

	优点	缺点
线下评估	无需真实用户参与；实验成本低	数据稀疏；与真实数据存在差距；不能代替真实满意度；无法评估商业指标
用户调研	风险低、允许收集定性数据，用户解释量化结果	执行代价昂贵（佣金、时间）实验设置要求严苛
线上评估	具有真实性	周期长；实验结果取决被实验的人

3 RS 评测指标

Oliver 提出的期望确认理论(Expectation Confirmation Theory，ECT)是研究消费者满意度的基础理论。用户在使用 RS 之前，会对系统推荐的物品产生某种期望，在使用物品或者体验服务之后，用户会对 RS 建立相应的感知，根据感知与期望的匹配程度确认对系统的满意度，进而决定是否再次使用该系统。本节从准确度指标与非准确度指标两个方面系统地论述各种 RS 的各类评测指标。

3.1 准确度指标

(1)评分准确度

在一些应用中，会有一个让用户打分的功能，例如在蚂蚁蜂窝上给一个旅游景点打分，当我们用推荐算法去预测用户对景点的评分(打几颗星)，此时关注的是评分的准确度，具体评估细则有平均绝对误差(MAE)，均方根误差(RMSE)，归一化平均绝度误差(NMAE)、归一化均方根误差(NRSE)。它们越小，则证明系统的准确度更高。其中，RMSE 使用平方根惩罚，结果更为严苛，NMAE、NRSME 是将 MAE、RMSE 归一化到(0,1)之间，使得归一化后偏差能在不同评分范围的不同应用之间可比。分别的定义如下：

$$MAE=\frac{\sum_{u\in U}\sum_{i\in test_u}\left|\hat{r}_{u,i}-r_{u,i}\right|}{\sum_{u\in U}\left|test_u\right|} \tag{1}$$

$$RMSE=\sqrt{\frac{\sum_{u\in U}\sum_{i\in test_u}\left(\hat{r}_{u,i}-r_{u,i}\right)^2}{\sum_{u\in U}\left|test_u\right|}} \tag{2}$$

$$NMAE=\frac{MAE}{r_{\max}-r_{\min}},NRMSE=\frac{RMSE}{r_{\max}-r_{\min}} \tag{3}$$

其中，U 为用户集合，$test_u$ 为测试集合，$\hat{r}_{u,i}r_{u,i}$ 分别代表预测评分与真实评分，$r_{\max}$ 和 $r_{\min}$ 分别代表用户评分中的上限值和下限值。值得注意的是，当用户仅在意推荐列表前端的误差，或者系统仅提供喜欢/不喜欢的选择信息，就不应采用评分准测度进行评估。

(2)排名准确度

多数含有 RS 的网站或应用上，通常将给用户展示的推荐设为列表。此时，我们关注的是列表的物品排序，此时，排序成为了衡量推荐结果的关键指标。排名指标中平均准确率(Mean Average Precision，*MAP*)和归一化折损积增益(Normalized Discounted Cumulative Gain，*NDCG*)是较为常用的指标。*MAP* 即 *AP*(准确度)的平均值。假设共有 M 个用户，则系统的 MAP 可以简单定义为：

$$MAP=\frac{\sum_{k=1}^{M}AP@n}{M} \tag{4}$$

其中，$AP@n=\sum_{i=1}^{n}(pred_i*cinr_i)$，$n$ 代表推荐列表的长度，$pred_i$ 表示前 i 个结果的正确率。$cinr_i$ 是一个二值项，如果第 i 个推荐错误，取值为 0。反之，取 $\frac{1}{n}$。

NDCG 是 *DCG* 的归一化，可以定义为：

$$NDCG=\frac{DCG}{DCG^*} \tag{5}$$

其中 *DCG* 在 *CG* 的基础上根据位置进行相关度加权，加权后的指标为：

$$DCG(P)=re_i+\sum_{i=1}^{P}\frac{2^{re_i}}{\log_b(i)} \tag{6}$$

re_i 表示第 i 位结果的得分，b 为自由参数，通常在 2 到 10 之间，但在使用时，一般令 $b=2$。

(3)分类准确度

分类准确度一般通过准确率和召回率、F 指标、AUC 等相关指标进行度量。对于任意用户 u，在推荐长度为 N 的列表中，则 u 的准确率与召回率表示如下：

$$Precision_u = \frac{\#tp}{\#tp + \#fp} \tag{7}$$

$$Recall_u = \frac{\#tp}{\#tp + \#fn} \tag{8}$$

其中 $\#tp$ 表示用户喜欢且系统成功推荐的物品的数量，$\#fn$ 表示用户喜欢但未被推荐的物品数量，$\#fp$ 代表用户不喜欢，但推荐给用户了的物品数量，$\#tn$ 代表了用户不喜欢，系统也没有推荐给用户的物品数量。$N = \#tp + \#fn$；系统整体的平均准确率、平均召回率则可以定义如下：

$$P(N) = \frac{\sum_u \mathrm{Precision}_u}{N} \tag{9}$$

$$R(N) = \frac{\sum_u \mathrm{Recall}_u}{N} \tag{10}$$

准确率和召回率是机器学习中通用的评估指标，但是其实它们其实在严格意义上来说是一对“矛盾”的评价指标。如果推荐集合的规模足够大，那么就可以观察到二者之间的关系，召回率提高，准确率会有所降低。M Pazzani 等人于 1997 年在文献[13]中首次提出 F-Measure 指标，同时考虑了准确率与召回率。该指标是准确率和召回率的调和平均值，用于综合反映系统的整体推荐效果，其可以定义为：

$$F = \frac{(\alpha^2 + 1)P(N) * R(N)}{\alpha^2(\mathrm{Precision} + R(N))} \tag{11}$$

除了 P(N)、R(N)与 F 指标外，常用的分类指标为 AUC 指标。AUC 指标表示 ROC 曲线下的面积，AUC 常用来处理二分类问题，如喜欢/不喜欢。ROC 曲线表征了推荐算法的整体表现，假设有三个算法 Algorithm0、Algorithm1、Algorithm2，它们的面积分别为 0.91，0.6,0.79，如图 1 所示，则算法 Algorithm0>Algorithm2>Algorithm1。(“>”表示“优于”)然而该指标没有考虑推荐位置排序的影响，如果算法有相同的 ACU 值，尽管客观上两个推荐算法有优劣之分，但是 AUC 指标却无法作出判断。

3.2 非准确度指标

(1)覆盖率

一些算法可能只针对具有大量数据的物品才有高质量的推荐，例如 CF。而一个好的 RS 不仅需要具备较高的准确度，而且需要较高的覆盖率[14]。覆盖率(Coverage)表示 RS 对物品长尾的发掘能力[15]，用以衡量系统是否能推荐到所有物品。假设现有用户集合为 $User$，系统给每个用户 u 提供 Top-N 列表为 $R(u)$，则其覆盖率可以计算为：

$$Coverage = \frac{\left|User_{u \in User} R(u)\right|}{|I|} \tag{12}$$

此外，信息熵、基尼系数(Gini Index)[2][15]两个指标通常也用于定义覆盖率。它们分别定义为：

$$H = -\sum_{i=1}^{n} p(i)\log p(i) \tag{13}$$

$$G = \frac{1}{n-1}\sum_{j=1}^{n}(2j - n - 1)p(i_j) \tag{14}$$

其中 i_j 是按照物品流行度 p 按从小到大的顺序进行排序的物品列表中的第 j 个物品。

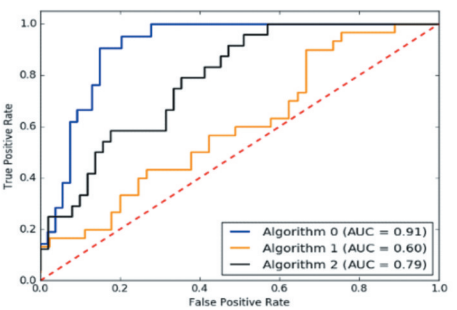

图1 ROC曲线示例

(2)多样性

用户的兴趣是广泛的，为了提高用户的感知质量，推荐结果应该能够覆盖用户的所有兴趣点，甚至包括用户未发觉的潜在兴趣，体现多样性。我们可将多样性简单理解为相似性的反面，在一些环境下，用户并不喜欢系统推荐的都是一系列相似的物品。例如某个用户刚刚已买了一个剃须刀，此时系统推荐其他品牌的剃须刀就显得没有多大的意义了。假设 $sim(i,j) \in [0,1]$ 为物品 i，j 的相似性。那么用户 u 的推荐列表 $R(u)$ 的多样性可定义为(15)。

$$Diversity(R(u)) = \frac{\sum_{i,j \in R(u), i \neq j}(1 - sim(i,j))}{\frac{1}{2}|R(u)|(|R(u)| - 1)} \tag{15}$$

那么整个 RS 的多样性可定义为(16)，其中 U 表示所有用户集合。

$$Diversity = \frac{1}{|U|}\sum_{u \in U} Diversity(R(u)) \tag{16}$$

以上讨论方法基于物品之间的相似性，在评价时

应保证在最小的影响精度为代价的前提下，提高推荐的多样性。

（3）新颖性与惊喜度

RS 中，“个性化”与“推荐”缺一不可，其中“个性化”是系统最大的特色。楼尊曾提出，稳定的心理特征中包括了“独特性需求”[16]。消费者自然也具有独特性需求，为体现其独特的个性，他们更倾向于购买个性化的商品。因此，在评估 RS 时，新颖度与惊喜度是两个重要的指标。新颖度指的是给用户推荐他并不知道的物品。最简单的方法即把与用户进行过交互的物品过滤掉，但这并不能完全体现新颖性，Celma Ò.曾研究了新颖性的评估，平均流行度是其中最为简便的方法。所谓的平均流行度即推荐列表中所有物品的度的总和的平均值。推荐冷门物品会比热门物品好，因为向用户推荐热门产品并没有达到个性化推荐的目的。换而言之，推荐列表中的物品平均流行度越小，对用户来说，新颖度就越高，效果越好。而惊喜度在近几年一直受到学术界的关注，它跟新颖度有些类似，但比新颖新要求更严格。若推荐的结果和用户的兴趣并不具备相似性，但是却能够让用户感到满意，此时，推荐具有较高的惊喜度。举个简单的例子，假设用户 u 没有看过电影 I，当系统将 I 推荐给 u 时，如果 I 与 u 的兴趣无关，u 自然会觉得很奇怪，但是，如果用户观看完电影觉得不错，那么可以说推荐是让用户惊喜的。

（4）健壮性

RS，是基于用户私人数据的软件应用，随着相关应用不断扩展，人们对于 RS 的依赖越来越强烈，RS 容易成为恶意用户为获取商业利益而作弊攻击的对象。因此，在衡量系统性能时需要考虑系统的抵御作弊的能力，即健壮性。推荐算法由一般都需要通过分析用户的行为实现，若用户行为是恶意的，推荐结果必将受到影响。创建一个能免除任何攻击的系统是不切实际的[17]，模拟攻击是评估系统健壮性的常用手段。学者在研究推荐算法时可以利用算法对用户先生成一个推荐列表，然后往数据集中注入一定的噪声数据，然后再次生成推荐列表，最后对比两次推荐列表的相似度，以此评估算法的健壮性。健壮性的另一方面指的是在极端条件下系统的稳定性，例如某一时间断大规模的用户请求。例如每年双十一，淘宝就有大规模用户请求。这种情况下，RS 的健壮性实际上与硬件规模与可靠性、数据库软件等基础设施相关。基于并行计算的推荐算法，例如在 Hadoop 上利用 MapReduce 实现的推荐算法通常具有相对较高的健壮性。

4 结语

RS 的社会意义是以“人”为本[18]，它指导用户在信息海洋中寻找灯塔，明确前行的方向。而其评估是检验系统性能手段，对 RS 研究中起到了引导作用。它是物的范畴，而其中的“个性化”则归属于人的范畴。在评估 RS 时需要进行综合考虑。传统的评估指标都存在固有的问题，并不能确切地反映系统的质量。网站重视系统带来的效益，而用户关心自身体验，因此如何设计一个在提升用户满意度的同时提高网站收入的推荐算法以及其评价指标体系，是一个很有研究意义的方向。若能解决评价指标存在的问题，定能推动 RS 的研究工作不断发展。本文对评估 RS 的指标进行了系统的综述，并认为可以从以下五个方面继续研究：

（1）保护用户隐私。在现有经典的推荐算法中，大多数是利用用户数据产生推荐的，如果要得到高准确率、高用户体验的推荐势必会更深层收集与挖掘用户数据，这就给用户隐私带来了一定的威胁。一个好的 RS，理应既提供准确、合理的推荐，又确保用户信息不被恶意用户随意获取。现有的文献中，考虑到用户隐私的推荐算法研究较少。因此在对系统进行评估时应考虑是否保护了用户隐私。

（2）推荐算法的时间效率。快节奏的时代“快”是一种趋势。推荐算法能否在短时间内处理爆炸量的信息，给出实时的主动推荐也是值得考虑的因素。

（3）算法的普适性。从复杂的指标群中选择合适的指标去评估系统极具挑战性。不同的推荐算法应用在不同的数据集中效果不同，目前未存在能够应用于所有领域的推荐算法。如果推荐算法在不同的数据集得到的效果相差不大，即说明该算法普适性强。

（4）添加客观性生理指标。评测是一个价值判断的过程，在用户调查研究中用户的作答是主观的，未必是用户真实的想法与感受。因此在评价中应引入一些客观性的指标辅助评估，例如，心率、血压、脑成像等，以提高结论的可信度。

（5）考虑移动的个性化 RS。《2016 年中国移动社交电商发展专题报告》中指出，手机网民的规模已高达

6.2 亿,并且逐年增加。现今移动 RS 的研究的学术文献并不少,但针对性的评价指标却很少,这是一个未来一个富有吸引力的研究方向。

参考文献:

[1]孙海峰,甘明鑫等. 国外电影 RS 网站研究与评述[J]. 计算机应用,2013,33(S2):119-124.

[2]Liang HU,et al. Improving the Quality of Recommendations for Users and Items in the Tail of Distribution[J]. ACM Transactions on Information Systems,2017,35(3):25.

[3]Kalamatianos G,et al. Suggesting Points of Interest via Content Based,Collaborative,and Hybrid Fusion Methods in Mobile Devices [J]. ACM Transactions on Information Systems,2017,36(3):23.

[4]Chenyi Zhang,et al. Trip Recommendation Meets Real-World Constraints:POI Availability,Diversity, and Traveling Time Uncertainty [J]. ACM Transactions on Information Systems, 2016,35(1):5.

[5]Da Cao,et al. Cross-Platform App Recommendation by Jointly Modeling Ratings and Texts[J]. ACM Transactions on Information Systems, 2017,35(4):37.

[6]Hao Ma et al. Improving Recommender Systems by Incorporating Social Contextual Information[J]. ACM Transactions on Information Systems, 2011, 29(2):9.

[7]Zhiyong Cheng,et al. On Effective Location-Aware Music Recommendation[J]. ACM Transactions on Information Systems,2016, 34(2): 1-32.

[8]Yong Ge et al. Cost-Aware Collaborative Filtering for Travel Tour Recommendations[J]. ACM Transactions on Information Systems,2014, 32(1):4.

[9]Loc Do,Hady W.Lauw. Probabilistic Models for Contextual Agreement in Preferences[J]. ACM Transactions on Information Systems, 2016,34-(4):21.

[10]Rana Forsati,et al. Matrix Factorization with Explicit Trust and Distrust Side Information for Improved Social Recommendation[J]. ACM Transactions on Information Systems (TOIS),2014,32(4):17.

[11]Shi L, Zhao W X, Shen Y D. Local Representative-Based Matrix Factorization for Cold-Start Recommendation[J]. ACM Transactions on Information Systems,2017,36(2):1-28.

[12]郭磊,马军,陈竹敏,姜浩然. 一种结合推荐对象间关联关系的社会化推荐算法[J]. 计算机学报,2014,37-(01):219-228.

[13]Pazzani M,Billsus D. Learning and Revising User Profiles:The Identification of Interesting Web Sites[J]. Machine Learning,1997,27(3): 313-331.

[14]项亮.推荐系统实践[M].北京:人民邮电出版社,2017:87-89.

[15]Shani G,Gunawardana A. Evaluating Recommendation Systems[J]. Recommender Systems Handbook, 2011:257-297.

[16]楼尊. 参与的乐趣——一个有中介的调节模型[J]. 管理科学,2010,69-76.

[17]Francesco Ricci. 推荐系统技术、评估及高效算法[M] .李艳民等译.北京:机械工业出版社,2015:178-190.

[18]刘凯,王伟军,黄英辉,方璐. 个性化 RS 理论探索:从系统向用户为中心的演进[J]. 情报理论与实践,2016,39(03):52-56.

作者简介:

刘春霞(1994-),女,广西防城港人,硕士,研究方向为推荐系统、智能计算

武玲梅(1990-),女,山西长治人,硕士,研究方向为机器学习、推荐系统

谢小红(1995-),女,广西北海人,硕士研究生,研究方向为新闻推荐、智能计算

收稿日期:2018-04-26　　修稿日期:2018-07-09

(下转第 20 页)

Prediction of Domestic Movie Box Office Based on BP Neural Network

LI Yi, WANG Xiao-feng

(College of Information Engineering, Shanghai Maritime Univeristy, Shanghai 201306)

Abstract:

In order to predict the film box office of domestic film before its release, selects 440 representative films made in China from 2010 to 2017 as samples. After reading a large number of literature at home and abroad, selects 20 influence factors from them. Analyzes and forecasts the 440 factors through the modeling of BP neural network. The experimental results show that most of the sample data errors are controlled within 10%, thus guide the risk control of film investment.

Keywords:

Box Office; Neural Network; Prediction

（上接第 15 页）

Research Review of Recommended System Assessment

LIU Chun-xia, WU Ling-mei, XIE Xiao-hong

（College of Computer and Information Engineering, Guangxi Teachers Education University, Nanning 530299）

Abstract:

The recommendation system is gradually popular in all major commercial fields and scientific research fields, and there are a lot of related researches. However, most scholars only focus on the research of the recommendation algorithm and neglect the research on the evaluation of the recommendation system. Evaluation is a means to test the performance of the recommender system and has important guiding significance for the development of the recommender system. Focuses on the evaluation of the research, conducts surveys, analyses, and summarizes relevant documents published in the famous magazine ACM TOIS in the past five years, and points out the current status of the evaluation research, explains the advantages and limitations of the three experiments. It also focuses on a systematic review of existing evaluation indicators. Based on the current problems in the evaluation study, proposes five future directions.

Keywords:

Information Overload; Recommendation Systems; Systematic Evaluation; Evaluation Index