推荐系统评价指标综述

朱郁筱[1]，吕琳媛[2,3,4]

(1. 电子科技大学互联网科学中心 成都 610054; 2. 杭州师范大学信息经济研究所 杭州 310036;
3. 弗里堡大学物理系 瑞士 弗里堡 CH-1700; 4. 百分点推荐技术研究中心 成都 611731)

【摘要】对现有的推荐系统评价指标进行了系统的回顾，总结了推荐系统评价指标的最新研究进展，从准确度、多样性、新颖性及覆盖率等方面进行多角度阐述，并对各自的优缺点以及适用环境进行了深入的分析。特别讨论了基于排序加权的指标，强调了推荐列表中商品排序对推荐评价的影响。最后对以用户体验为中心的推荐系统进行了详细的讨论，并指出了一些可能的发展方向。

关 键 词 评价指标; 海量信息; 信息系统; 推荐系统
中图分类号 TP391 **文献标识码** A doi:10.3969/j.issn.1001-0548.2012.02.001

Evaluation Metrics for Recommender Systems

ZHU Yu-xiao[1] and LÜ Lin-yuan[2,3,4]

(1. Web Science Center, University of Electronic Science and Technolgy of China Chengdu 610054;
2. Institution of Information Economy, HangZhou Normal University Hangzhou 310036;
3. Department of Physics, University of Fribourg Fribourg Switzerland CH-1700;
4. Baifendian Research Center on Recommender Systems Chengdu 611731)

Abstract In this article, the existed evaluation metrics for recommender systems are reviewed and the new progresses in this field are summarized from four aspects: accuracy, diversity, novelty and coverage. The merits, weaknesses and applicable conditions of different evaluation metrics are analized. The focus is concentrated on the importance of rank and some representative rank-sensitive metrics. The user-centric recommender systems are discussed and some important open problems are outlined as future possible directions.

Key words evaluation metrics; huge information; information systems; recommender systems

Web 2.0网络技术和社会化媒体的大力发展使得每个人既可以是信息的接收者也可以是信息的创造者。通过互联网人们可以以更快速、便捷、低成本的方式获取所需的信息。然而，在面对一个丰富多彩的网络世界的同时也面临着信息过载的问题[1-5]。如果说一百年前全世界的信息总量装满了西湖，如今我们面对的是整个太平洋[6]。虽然网络带来了更多的选择，但数量庞大及自身质量差异使得如何从这些海量的信息中识别出真正有价值的信息变得越来越困难。要想征服这个浩瀚的海洋就要打造一艘博流的方舟——大力发展先进的信息过滤技术。推荐系统应运而生，它被认为是解决信息过载问题的一个有效的方式。与传统的信息过滤技术搜索引擎不同，推荐系统不需要用户提供用于搜索的关键词，它将通过分析用户的历史交易记录或行为挖掘用户的潜在兴趣，进而对其进行推荐。因此，推荐系统更能满足用户个性化的需求。个性化推荐系统在电子商务网站上已经得到了广泛的应用，并带来了巨大的商业价值。

目前的推荐算法主要包括协同过滤算法[7]、基于内容的推荐算法[8]、谱分析[9-10]、基于扩散的方法[11-14]以及混合推荐算法[15-22]。文献[23]对已有的方法进行了全面的总结和分析，特别强调了物理方法在推荐系统中的应用，并在同一组数据集上进行了比较，是目前在推荐系统研究领域发表的最全面的综述。此外，基于这些基本方法的改进方法层出不穷。面对众多的推荐算法，如何有效客观地评价推荐系统的优劣却是颇具挑战的问题[24-26]。虽然目前已有的推荐系统评价指标数不胜数，但是国内外学者对此问题认识仍然不足，主要表现在3个方面。首

收稿日期：2012－01－18
基金项目：瑞士国家科学基金(200020-132253)
作者简介：朱郁筱(1988－)，女，博士生，主要从事信息物理，包括链路预测、推荐系统以及传播动力学等方面的研究

先，很多学者对推荐评价指标认识不全面，有些学者只局限于推荐的精确性一个方面，而对多样性、新颖性、覆盖率等指标视而不见。其次，由于学术界没有建立推荐算法评估完整统一的指标群，部分学者在撰写论文的时候倾向于选择对自己算法有利的指标，而对其他指标的表现只字不提，本文在重现一些文献工作的时候发现，某些学者所提出的算法在他们没有涉及的指标上面往往表现很差。最后，一些学者对于各个指标所能衡量的算法的性能方面以及不同指标的优劣和适用性了解较少，因此在评价指标的选择和结果解释方面存在不足之处。综上所述，本文认为，客观合理地评价指标体系的建立会极大地促进推荐算法的研究和推荐系统的开发——希望对此有所贡献。

本文对现有推荐系统评价指标进行了系统的回顾，总结了推荐系统评价指标的最新研究进展，从准确度、多样性、新颖性及覆盖率等多角度进行阐述，并对各自的优缺点以及适用环境进行了深入分析。特别讨论了基于排序加权的指标，强调了推荐列表中商品排序对推荐评价的影响。另外对在线和离线的测试加以区分。

1 测评方法

推荐系统的评价可分为在线评价和离线评价两种方式。在线评价其实就是设计在线用户实验，根据用户在线实时反馈或事后问卷调查等结果来衡量推荐系统的表现。目前最常用的在线测试方法之一是*A*/*B*测试，所谓*A*/*B*测试，简单说来，就是为了同一个目标制定两个方案，让一部分用户使用*A*方案，另一部分用户使用*B*方案，记录下用户的使用情况，看哪个方案更符合设计目标。它的核心思想是：1) 多个方案并行测试；2) 每个方案只有一个变量不同；3) 以某种规则优胜劣汰。其中第2点暗示了*A*/*B*测试的应用范围：*A*/*B*测试必须是单变量。待测试方案有非常大的差异时一般不太适合做*A*/*B*测试，因为它们的变量太多了，变量之间会有很多的干扰，所以很难通过*A*/*B*测试的方法找出各个变量对结果的影响程度。显然，*A*/*B*测试用于在推荐系统的评价中就对应于唯一的变量——推荐算法。注意，虽然*A*/*B*测试名字中只包含*A*、*B*，但并不是说它只能用于比较两个方案的好坏，事实上，完全可以设计多个方案进行测试，*A*/*B*测试这个名字只是习惯的叫法而已。不同的用户在一次浏览过程中，看到的应该一直是同一个方案。如他开始看到的是*A*方案，则在此次会话中应该一直向他展示*A*方案，而不能一会儿让他看*A*方案，一会儿让他看*B*方案。同时，还需要注意控制访问各个版本的人数，大多数情况下希望将访问者平均分配到各个不同的版本上。

这种在线测试方式虽然可以直观地得到用户对系统的满意度等指标，但是从设计实验到施行实验整个过程所需的高额成本却是一般的科研工作者都无法负担的。所以目前的大部分研究都集中于离线测评上。所谓的离线测评也即是根据待评价的推荐系统在实验数据集上的表现，然后再根据下文将要提到的评价指标来衡量推荐系统的质量。相对于在线评价，离线评价方法更方便更经济，一旦数据集选定，只需要将待评测的推荐系统在此数据集上运行即可。但是离线评价也面临着以下问题：

1) 数据集的稀疏性限制了适用范围，例如不能用一个不包含某用户任何历史记录的数据集来评价推荐系统对该用户的推荐结果。

2) 评价结果的客观性，由于用户的主观性，不管离线评测的结果如何好，都不能得出用户是否喜欢某推荐系统的结论。

3) 难以找到离线评价指标和在线真实反馈(如点击率、转化率、点击深度、购买客单价、购买商品类别等)之间的关联关系。

尽管如此，在目前的研究工作中离线评价方式仍是科研工作人员的首选。离线评价方式最主要的两个环节就是数据集的划分以及评价指标的选择。目前最为常用的数据划分方式仍为随机划分。推荐系统所用的数据集为用户和商品的二元关系信息，可以用一个用户-商品的二部分图表示，即$G(U,O,E)$，其中，U表示用户集合，O表示商品集合，E在不同的系统中具有不同的含义。在有评分系统中它表示explicit data, 也即是用户评分集合，包括正分和负分。在常见的“顶-踩”系统中表示binary data，即只有用户喜欢不喜欢的信息，并没有涉及具体评分。在另外一类系统中，它代表unary data，如只是用户购买或者浏览的数据信息，这里并没有包含用户对商品真实喜好的数据。目前对于此类系统，最常用的做法是假设用户的购买或者浏览行为就代表了他的喜好，这难免有些牵强，针对这一问题，本文后面也有相关讨论。定义$M=|U|$为系统中的用户数量，$N=|O|$为系统中商品数量。所谓的随机划分数据集，就是在集合E中随机选取一定比例作为测试集E^{P}，剩下的部分就是训练集E^{T}，显然$E=E^{\mathrm{P}}+E^{\mathrm{T}}$，$E^{\mathrm{P}}\cap E^{\mathrm{T}}=\varnothing$。定义$E_u^{\mathrm{P}}$为测试集中与用户$u$相关的商品集合，$E_u^{\mathrm{T}}$为训练集中与用户$u$相关的商品集合。

离线评价就是将训练集的信息作为算法的输入进行推荐，然后将推荐结果和测试集的信息进行比较并利用已有的评价指标来衡量推荐系统的表现。本质上推荐算法可以看成是在二部分图上的链路预测问题[27-29]。值得注意的是，完全的随机划分可能会导致有些用户或者商品没有被划分进测试集，或者划分后变成孤立商品。为了避免这类情况发生，在实际操作时往往针对每个用户划分一定比例的数据作为测试集。

2　准确度指标

推荐的准确度是评价推荐算法最基本的指标。它衡量的是推荐算法在多大程度上能够准确预测用户对推荐商品的喜欢程度。目前大部分的关于推荐系统评价指标的研究都是针对推荐准确度的。准确度指标有很多种，有些衡量的是用户对商品的预测评分与真实评分的接近度，有些衡量的是用户对商品预测评分与真实评分的相关性，有些考虑的是具体的评分，有些仅仅考虑推荐的排名。本文将准确度指标分为4类，即预测评分准确度、预测评分关联性、分类准确度和排序准确度，下面将对每一类进行详细介绍。

2.1　预测评分的准确度

顾名思义，预测评分的准确度衡量的是算法预测的评分和用户的实际评分的贴近程度。这个指标在需要向用户展示预测评分的系统中尤为重要。如MovieLens的电影推荐系统[5]就是预测用户会对电影打几颗星，一颗星表示很糟糕的电影，五颗星则表示不得不看的电影。值得注意的是，即便一个推荐算法能够比较成功的预测出用户对其他商品的喜好排序，但它在评分准确度上的表现仍然可能不尽人意，这也是商业领域的大部分推荐系统只向用户提供推荐列表而没有预测评分的主要原因。

预测评分的准确度指标目前有很多，这类指标的思路大都很简单，就是计算预测评分和真实评分的差异。最经典的是平均绝对误差(mean absolute error，MAE)[30-32]，如果用 $r_{u\alpha}$ 表示用户 u 对商品 α 的真实评分，$r'_{u\alpha}$ 表示用户 u 对商品 α 的预测评分，E^{P} 表示测试集，那么 MAE 定义为：

$$\mathrm{MAE}=\frac{1}{\left|E^{\mathrm{P}}\right|}\sum_{(u,\alpha)\in E^{P}}\left|r_{u\alpha}-r'_{u\alpha}\right| \tag{1}$$

MAE因其计算简单、通俗易懂得到了广泛的应用。不过MAE指标也有一定的局限性，因为对MAE指标贡献比较大的往往是那种很难预测准确的低分商品，所以即便推荐系统*A*的MAE值低于系统*B*，很可能只是由于系统*A*更擅长预测这部分低分商品的评分，也即是系统*A*比系统*B*能更好的区分用户非常讨厌和一般讨厌的商品罢了，显然这样的区分意义并不大。

除了计算所有预测商品的平均绝对误差外，文献[30]曾主张只考虑用户比较敏感的商品的预测误差。如在一个7分制的系统中，根据用户的评分将所有的商品分为3类，其中评分大于5和小于3的商品被看作是用户比较敏感的，认为用户主要关注推荐系统在比较敏感的商品上的表现。

此外，平均平方误差(mean squared error，MSE)、均方根误差(root mean squared error，RMSE)以及标准平均绝对误差[15](normalized mean absolute error，NMAE)都是与平均绝对误差类似的指标。它们分别定义为：

$$\mathrm{MSE}=\frac{1}{|E^{P}|}\sum_{(u,\alpha)\in E^{P}}(r_{u\alpha}-r'_{u\alpha})^{2} \tag{2}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{|E^{P}|}\sum_{(u,\alpha)\in E^{P}}(r_{u\alpha}-r'_{u\alpha})^{2}} \tag{3}$$

$$\mathrm{NMAE}=\frac{\mathrm{MAE}}{r_{\max}-r_{\min}} \tag{4}$$

式中，$r_{\max}$ 和 $r_{\min}$ 分别为用户评分区间的最大值和最小值。由于MSE和RMSE指标对每个绝对误差首先做了平方，所以这两个指标对比较大的绝对误差有更重的惩罚。NMAE由于在评分区间上做了归一化，从而可以在不同的数据集上对同一个推荐算法表现进行比较。

2.2　预测评分关联

预测评分关联是用来衡量预测评分和用户真实评分之间的相关性的，最常见的3种相关性指标分别是Pearson积距相关[33]、Spearman相关[34]和Kendall's Tau[35]。Pearson积距相关系数衡量的是预测评分和真实评分的线性相关程度，定义为：

$$\mathrm{PCC}=\frac{\sum_{\alpha}(r'_{\alpha}-\bar{r}')(r_{\alpha}-\bar{r}')}{\sqrt{\sum_{\alpha}(r'_{\alpha}-\bar{r}')^{2}}\sqrt{\sum_{\alpha}(r_{\alpha}-\bar{r}')^{2}}} \tag{5}$$

式中，r_{α} 和 r'_{α} 分别表示商品 α 的真实评分和预测评分。Spearman关联和Pearson关联定义的形式是一样的，唯一不同的是Spearman关联考虑的不是预测评分值，而是根据预测评分值所得到的排序值，即将式(5)中的 r_{α} 和 r'_{α} 分别替换成商品 α 的真实排名和预测排名。

与Spearman类似，Kendall's Tau也是刻画两种排

序值的统一程度的，它定义为：

$$\tau = \frac{C - D}{C + D} \tag{6}$$

式中，C为正序对的数目；D表示逆序对的数目。显然当所有的商品对都是正序对时$\tau = 1$，当所有的商品对都是逆序对时$\tau = -1$。如某用户对商品1～5的真实排序名依次为：O_1、O_2、O_3、O_4、O_5，对应于10种序关系，即$(O_1>O_2)(O_1>O_3)(O_1>O_4)(O_1>O_5)(O_2>O_3)(O_2>O_4)(O_2>O_5)(O_3>O_4)(O_3>O_5)(O_4>O_5)$，而推荐系统$A$的预测排序为：$O_2$、$O_1$、$O_4$、$O_5$、$O_3$，对应于10种序关系$(O_1<O_2)(O_1>O_3)(O_1>O_4)(O_1>O_5)(O_2>O_3)(O_2>O_4)(O_2>O_5)(O_3<O_4)(O_3<O_5)(O_4>O_5)$，对比上述两组序关系得到3个逆序对，即$(O_1,O_2)(O_3,O_4)(O_3,O_5)$，剩余7个为正序对，根据式(6)得到推荐系统$A$的$\tau$值为0.4。

预测评分关联指标尽管计算简单，但是却没有在推荐系统评价中被广泛采纳。因为以上指标都有一定的不足之处。以Kendall's Tau为例，它对于所有的排名偏差都分配相等的权重，而不管具体的排序值。但是不同的排名偏差显然是不能同等对待的，如某用户对于100个商品的真实排名是1、2、…、98、99、100，推荐系统A得出的预测排名是2、1、3、4、…、99、100，而推荐系统B得出的预测排名是1、2、…、98、100、99，根据式(6)易得系统A和系统B的τ值相等，但是实际上推荐系统B可能更好些，因为用户往往更关心排在前面的推荐商品，因此对排在前面的商品推荐的准确度也更为敏感。

另外，在实际系统中可能有某用户对某两个或者两个以上的商品评分一致的情况，也即是所谓的弱关系排序问题。显然以上所提及的预测评分关联指标都不适用于此种情形。当真实排名或预测排名有并列情况出现时，可以用基于Kendall's Tau改进的一个指标[24]来衡量，定义为：

$$\tau \approx \frac{C - D}{\sqrt{(C + D + S_{\mathrm{P}})(C + D + S_{\mathrm{T}})}} \tag{7}$$

式中，S_{T}表示真实评分相同的商品对数量；S_{P}表示预测评分相同的商品对数量。

此外，为了比较两个不同的弱排序序列，文献[36]提出了一种归一化的基于距离的评价指标(normalized distance-based performance measure，NDPM)。它的主要思想是先统计两个排序相悖的商品对个数C^-以及两个排序兼容的商品对个数C^u。排序相悖是指在两个商品α和β中系统预测的是某用户更喜欢商品α，然而实际上用户更喜欢是商品β。排序兼容指的是系统预测用户对商品α和β同等喜欢，然而实际上用户更喜欢的是商品α或者是商品β。假如用T表示用户实际评分中具有严格偏好差别的商品对个数，则NDMP指标定义为：

$$\mathrm{NDMP} = \frac{2C^- + C^u}{2T} \tag{8}$$

显然NDMP取值介于0和1之间，而且NDMP值越小，预测评分关联越大，也即是系统的预测结果越好。如假设某用户对商品1-5的实际评分依次分别为：4.4、3.9、3.8、3.9、1.0，而某推荐系统预测的该用户对以上商品的评分依次为：3.3、3.1、3.0、4.3、3.1，那么此时具有严格偏好差别的商品对有(1,2)、(1,3)、(1,4)、(1,5)、(2,3)、(2,5)、(3,4)、(3,5)、(4,5)，其中排序相悖的商品对有(1,4)、(3,5)，排序兼容的商品对有(2,5)，由式(8)易得该系统的NDMP值为0.278。相对于其他的预测评分关联性指标，NDMP指标有一定的优势，它不仅适用于弱关系排序问题还可以用来评价推荐算法在不同数据集上的表现。值得注意的是，这些预测评分关联性指标都是只关注于预测排序值而不关注具体的预测评分值，所以它们都不适用于那些旨在为用户提供精确预测评分值的系统。

2.3 分类准确度

分类准确度指标衡量的是推荐系统能够正确预测用户喜欢或者不喜欢某个商品的能力。它特别适用于那些有明确二分喜好的用户系统，即要么喜欢要么就不喜欢。对于有些非二分喜好系统，在使用分类准确度指标进行评价的时候往往需要设定评分阈值来区分用户的喜好。如在5分制系统中，通常将评分大于3的商品认为是用户喜欢，反之认为用户不喜欢。如Yahoo的音乐推荐系统中，★表示再也不会听，★★表示平庸之作，★★★表示比较好听，★★★★表示很好听，★★★★★表示非常好听。与预测评分准确度不同的是，分类准确度指标并不是直接衡量算法预测具体评分值的能力，只要是没有影响商品分类的评分偏差都是被允许的。目前最常用的分类准确度指标有准确率(precision)、召回率(recall)、F_1指标和AUC。

文献[38-41]最先把准确率和召回率引入到推荐系统评价中。准确率表示用户对系统推荐商品感兴趣的概率。在计算准确率的时候，常用的做法是设定推荐列表长度L，根据预测评分对所有待预测商品排序，系统认为排在前L位的商品是用户最可能喜欢的，因此将它们推荐给用户。于是，对于一个未曾被用户选择或评分的商品，最终可能的结果有4种，即系统推荐给用户且用户很喜欢，系统推荐给用户

但是用户不喜欢，用户喜欢但是系统没有推荐，用户不喜欢且系统没有推荐。表1总结了这4种可能的情况，其中 N_{tp}， N_{fn}， N_{fp} 和 N_{tn} 分别表示4种情况的数目。B_u 表示用户 u 喜欢的商品数，显然 $L=N_{tp}+N_{fp}$， $B_u=N_{tp}+N_{fn}$。

表1　待预测的商品可能的4种情况

用户喜好	系统推荐	系统不推荐
喜欢	Ture-Positive N_{tp}	False-Negative N_{fn}
不喜欢	False-Positive N_{fp}	Ture-Negative N_{tn}

对于某一用户u，其推荐准确率为系统推荐的L个商品中用户喜欢的商品所占的比例，即：

$$P_u(L)=\frac{N_{tp}}{L}=\frac{N_{tp}}{N_{tp}+N_{fp}} \tag{9}$$

在离线测试中，N_{tp} 的值就等于同时出现在用户u的测试集合和其推荐列表中的商品的数目。在在线测试的时候，N_{tp} 的值将根据用户的实际反馈结果进行统计得到。将系统中所有用户的准确率求平均得到系统整体的推荐准确率，即：

$$P(L)=\frac{1}{M}\sum_u P_u(L) \tag{10}$$

式中，M表示测试用户的数量。注意，如果不是对系统的所有用户都进行考察，那么M值将小于系统中实际用户的数目。如有些基于网络随机游走的推荐算法[13,50,53]为保证测试网络的连通性往往不会抽取度为1的用户作为测试用户。使用这种方式获得的系统准确率保证了每个用户的贡献是平权的。在线下测试中，准确率会受评分稀疏性的影响。例如系统对一个只给很少部分的商品打过分的用户的推荐准确率往往很低。但这并不能说明推荐系统的效果很差，因为很有可能系统推荐的商品中有很多是用户没有打过分但是确实很喜欢的商品。在这种情况下，在线的测试结果，即用户的真实反馈，更能够准确地反应推荐系统的表现。

召回率表示一个用户喜欢的商品被推荐的概率，定义为推荐列表中用户喜欢的商品与系统中用户喜欢的所有商品的比率。对于用户u，其召回率为：

$$R_u(L)=\frac{N_{tp}}{B_u}=\frac{N_{tp}}{N_{tp}+N_{fn}} \tag{11}$$

在离线测试中，B_u 实际上就等于测试集中用户u喜欢的商品数，即$|E_u^P|$[40]。在实际应用中，由于不能准确知道系统没有推荐的商品中哪些是用户喜欢的，因此召回率很难应用于在线评估。将系统中所有用户的召回率求平均得到系统整体的召回率

$$R(L)=\frac{1}{M}\sum_u R_u(L) \tag{12}$$

在上述准确率和召回率的基础上，文献[13]将系统的推荐准确率和召回率与随机推荐的结果进行比较，首次提出了准确率提高比例和召回率提高比例的概念，定义为：

$$e_P(L)=P(L)\frac{MN}{B} \tag{13}$$

$$e_R(L)=R(L)\frac{N}{L} \tag{14}$$

式中，M和N分别代表系统中用户和商品的总个数；B表示用户喜欢的商品的总数。

推荐系统是根据商品满足用户喜好的可能性来进行推荐的，但是只有用户本人才知道商品是否符合自己口味，因此“喜欢”的界定在推荐系统是非常主观、因人而异的。不仅用户的兴趣有差异，用户的评分尺度也有差异的。如对于用户张三来说，评分3.5以上就说明他很喜欢这个商品了，但是对于比较苛刻的李四来说，可能2.5以上就代表很喜欢了。针对以上局限性，文献[16]提出了根据用户以往的评分历史来确定用户喜好评分阈值的方法。由于受到推荐列表长度，评分稀疏性以及喜好阀值等多方面因素的影响，很多学者不提倡利用准确率和召回率来评价推荐系统，特别是只单独考虑一种指标的时候误差极大。严格意义上，召回率是不适用于评价推荐系统的，因为它无形中已经假设了用户没有评分的商品都是他不喜欢的，这个假设显然是不合理的。

一种常用的方法是同时考虑准确率和召回率从而比较全面地评价算法的优劣。准确率和召回率指标往往是负相关的而且依赖于推荐列表长度[43]。一般情况下，随着推荐列表长度的增大，准确率指标会减小而召回率会增大。所以当一个系统没有固定的推荐列表长度时，就需要一个包含准确率和召回率的二维向量来反映系统的表现。为了方便起见，文献[44-45]提出 F_1 了指标，定义为：

$$F_1(L)=\frac{2P(L)R(L)}{P(L)+R(L)} \tag{15}$$

另外还有一些学者将准确率和召回率结合起来衡量信息检索结果的有效程度，如Average Precision、Precision-at-Depth、R-Precision、Reciprocal Rank[46]以及Binary Preference Measure[47]，对每种指标的具体定义参见文献[48]。

上述的一系列指标对于没有二分喜好的系统都是不太适用的。即给定一个推荐列表，当推荐的阀值不确定的时候，上述指标不再适用。在这种情况下，往往采用AUC指标[49]来衡量推荐效果的准确性。由于不受推荐列表长度和喜好阀值的影响，AUC指标被广泛应用于评价推荐系统中。

AUC指标表示ROC(receiver operator curve)曲线[42]下的面积，它衡量一个推荐系统能够在多大程度上将用户喜欢的商品与不喜欢的商品区分出来[49]。绘制ROC曲线的步骤如下：

1) 根据某一推荐算法产生一个商品推荐列表，即按照预测评分从高到低将待预测商品($\in U-E^{\mathrm{T}}$)排序。

2) 绘制ROC曲线坐标轴。横坐标为不相关的比例(percentage of non-relevant item)，纵坐标为相关的比例(percentage of relevant item)。横纵坐标的总长度都为1，横坐标的一单位长度为1除以不相关商品的数目，纵坐标的一单位长度等于1除以相关商品的数目。

3) 绘制ROC曲线从坐标点(0,0)开始，从排序列表第一位开始查看每一个商品是否符合下列3种情况中的一种。

(1) 如果该商品相关(如用户喜欢的商品)，则沿着*y*轴方向向上移动一单位。

(2) 如果该商品不相关(如用户不喜欢的商品)，则沿着*x*轴方向向右移动一单位。

(3) 如果该商品不确定其相关性，则舍弃该商品，不做任何移动。

注意，在仅有选择行为的系统中，如只有购买等交易行为的系统中，通常将测试集的商品看成是相关商品，将集合O-E中的商品看成是不相关的。而在评分系统中，有时候也将用户没有评分的商品(即不知道是否相关的商品)看成是不相关商品。如果推荐系统将所有用户喜欢的商品都排在不喜欢商品的前面，那么ROC曲线将是一条沿*y*轴竖直向上直到*y*=1的位置然后水平向右直到*x*=1的位置的折线，此时ROC曲线下面的面积为1，即AUC=1，对应于最完美的推荐。随机推荐的ROC曲线则大致对应于从原点(0,0)到(1,1)的对角线，此时ROC曲线下的面积为0.5，即AUC=0.5。关于ROC曲线的详细讨论可参见文献[24]。

由于ROC曲线绘制步骤比较繁琐，可以用以下方法来近似计算系统的AUC：每次随机从相关商品集，即用户喜欢的商品集中选取一个商品($\alpha \in E^{p}$)与随机选择的不相关商品($\beta \in O-E$)进行比较，如果商品α的预测评分值大于商品β的评分，那么就加一分，如果两个评分值相等就加0.5分。这样独立地比较n次，如果有n'次商品α的预测评分值大于商品β的评分，有n''次两评分值相等，那么AUC就可以近似写作：

$$\mathrm{AUC}=\frac{n'+0.5n''}{n} \tag{16}$$

显然，如果所有预测评分都是随机产生的，那么AUC=0.5。因此AUC大于0.5的程度衡量了算法在多大程度上比随机推荐的方法精确。AUC指标仅用一个数值就表征了推荐算法的整体表现，而且它涵盖了所有不同推荐列表长度的表现。但是AUC指标没有考虑具体排序位置的影响，导致在ROC曲线面积相同的情况下很难比较算法好坏，所以它的适用范围也受到了一些限制。

2.4 排序准确度

对于推荐排序要求严格的推荐系统而言，如果用评分准确度、评分相关性或者是分类准确度等指标来评价此类系统的好坏显然是不合适的。这类系统需要用排序准确度指标来度量算法得到的有序推荐列表和用户对商品排序的统一程度。排序准确度对于只注重分类准确度的系统来说太敏感了，它更适合于需要给用户提供一个排序列表的系统。如在比较两个推荐算法的时候，两个算法在推荐的5个商品中都有1个是用户感兴趣的，于是他们的推荐精确性都为0.2。但是算法*A*将用户喜欢的商品排在第1位，而算法*B*将用户喜欢的商品排在第5位，显然算法*A*更优越。考虑排序位置的影响，文献[50]提出了平均排序分(average rank score)来度量推荐系统的排序准确度。对于某一用户*u*来说，商品α的排序分定义如下：

$$\mathrm{RS}_{u\alpha}=\frac{l_{u\alpha}}{L_u} \tag{17}$$

式中，L_u表示用户*u*的待排序商品个数。在离线测试中L_u等于$|O-E_u^{\mathrm{T}}|$，也即用户*u*在测试集中的商品数目($|E_u^{\mathrm{p}}|$)加上未选择过的商品数目($|O-E_u|$)。$l_{u\alpha}$为待预测商品α在用户*u*的推荐列表中的排名(此时推荐列表长度为L_u)。离线测试中，L_u就等于用户*u*未选择过的商品数目。举例来说，如果有1 000部影片是用户*u*没有选择过的，其中用户喜欢的电影《金陵十三钗》出现在用户*u*推荐列表的第10位，那么对于用户*u*而言电影《金陵十三钗》的排序分为$\mathrm{RS}_{u\alpha}=10/1\,000=0.01$。将所有用户的排序分求平均即得到系统的排序分RS。排序分值越小，说明系统越趋向于把用户喜欢的商品排在前面。反之，则

说明系统把用户喜欢的商品排在了后面。由于平均排序分不需要额外的参数，而且不需要事先知道用户对商品的具体评分值，因此可以很好的比较不同算法在同一数据集上的表现。值得注意的是，在系统尺度(训练集和测试集)足够大的情况下，有 $\mathrm{AUC}+\mathrm{RS}\approx 1$。

3 基于排序加权的指标

现实生活中用户的耐心往往是有限的，一个人不太可能会不厌其烦地检查推荐列表中的所有商品，所以用户体验的满意度往往会受到用户喜欢的商品在推荐列表中位置的影响，这里介绍3个具有代表性的评价指标，更详细的信息参见文献[48]。

半衰期效用指标(half-life utility)[51]是在用户浏览商品的概率与该商品在推荐列表中的具体排序值呈指数递减的假设下提出的，它度量的是推荐系统对一个用户的实用性也即是用户真实评分和系统默认评分值的差别。用户 u 的期望效用定义为：

$$\mathrm{HL}_u=\sum_{\alpha}\frac{\max(r_{u\alpha}-d,0)}{2^{(l_{u\alpha}-1)/(h-1)}} \tag{18}$$

式中，$r_{u\alpha}$ 表示用户u对商品 α 的实际评分；而 $l_{u\alpha}$ 为商品 α 在用户u的推荐列表中的排名；d 为默认评分(如说平均评分值)；h 为系统的半衰期，也即是有50%的概率用户会浏览的推荐列表的位置。显然，当用户喜欢的商品都被放在推荐列表的前面时，该用户的半衰期效用指标达到最大值。系统的半衰期效用值定义为：

$$\mathrm{HL}=100\frac{\sum_u \mathrm{HL}_u}{\sum_u \mathrm{HL}_u^{\max}} \tag{19}$$

式中，$\mathrm{HL}_u^{\max}$ 为用户u的期望效用能达到的最大值。目前半衰期效用指标的使用仍然是有很大的局限性。首先参数的选取尚未有统一的标准，不同的学者可能会选择不同的参数，这样难免会造成混乱。另外，用户的浏览概率与商品在推荐列表中的位置呈指数递减这一假设并不是在所有系统中都适用。

折扣累计利润[52](discounted cumulative gain，DCG)的主要思想是用户喜欢的商品被排在推荐列表前面比排在后面会更大程度上增加用户体验，定义为：

$$\mathrm{DCG}(b,L)=\sum_{i=1}^{b} r_i+\sum_{i=b+1}^{L}\frac{r_i}{\log_b i} \tag{20}$$

式中，r_i 表示排在第 i 位的商品是否是用户喜欢的；$r_i=1$ 表示用户喜欢该商品；$r_i=0$ 表示用户不喜欢该商品；b是自由参数多设为2；L为推荐列表长度。

与DCG指标不同，排序偏差准确率[48](rank-biased precision，RBP)假设用户往往先浏览排在推荐列表首位的商品然后依次以固定的概率p浏览下一个，以1–p的概率不再看此推荐列表。RBP定义为：

$$\mathrm{RBP}(p,L)=(1-p)\sum_{i=1}^{L} r_i p^{i-1} \tag{21}$$

RBP和DCG指标的唯一不同点在于RBP把推荐列表中商品的浏览概率按等比数列递减，而DCG则是按照log调和级数形式。

4 覆盖率

覆盖率指标[24]是指算法向用户推荐的商品能覆盖全部商品的比例，如果一个推荐系统的覆盖率比较低，那么这个系统很可能会由于其推荐范围的局限性而降低用户的满意度，因为低的覆盖率意味着用户可选择的商品很少。覆盖率尤其适用于那些需要为用户找出所有感兴趣的商品的系统。覆盖率可以分为预测覆盖率(prediction coverage)，推荐覆盖率(recommendation coverage)和种类覆盖(catalog coverage)3种。

预测覆盖率表示系统可以预测评分的商品占所有商品的比例，定义为：

$$\mathrm{COV}_P=\frac{N_d}{N} \tag{22}$$

式中，N_d 表示系统可以预测评分的商品数目；N 为所有商品数目。

推荐覆盖率[53]表示系统能够为用户推荐的商品占所有商品的比例，显然这个指标与推荐列表的长度 L 相关。其定义为：

$$\mathrm{COV}_R(L)=\frac{N_d(L)}{N} \tag{23}$$

式中，$N_d(L)$ 表示所有用户推荐列表中出现过的不相同的商品的个数。推荐覆盖率越高，系统给用户推荐的商品种类就越多，推荐多样新颖的可能性就越大。如果一个推荐算法总是推荐给用户流行的商品，那么它的覆盖率往往很低，通常也是多样性和新颖性都很低的推荐。

种类覆盖率(COV_C)表示推荐系统为用户推荐的商品种类占全部种类的比例。相比预测覆盖率和推荐覆盖率，种类覆盖率应用的还很少。在计算种类覆盖率的时候，需要事先对商品进行分类。仅用覆盖率来衡量推荐系统的表现是没有意义的，它需要和预测准确度一起考虑[54]。如系统中某个类别的商品所有的用户都不喜欢，那么一个好的推荐算法可能再也不会向用户推荐这类商品，这时候它的覆盖率可能很低，但是准确度还是很高的。一个好的推

荐系统应在保证推荐准确度的同时尽量提高覆盖率。

目前评价覆盖率的指标还不够成熟，我们期待广大学者能够设计出一种更普适更合理的推荐系统覆盖率评价指标，这个评价指标应该符合以下几条标准：1) 能够同时考虑预测覆盖率(或推荐覆盖率)以及种类覆盖率；2) 对于用户比较喜欢的商品应赋予较高的权重；3) 能够在一定程度上结合预测准确率指标，从而减少评价的片面性。

5 多样性和新颖性

实际应用中，已经发现即使是准确率比较高的推荐系统也不能保证用户对其推荐结果满意[24]。一个好的推荐系统应该向用户推荐准确率高并且又有用的商品。譬如，系统推荐了非常流行的商品给用户，虽然可能使得推荐准确度非常高，但是对于这些信息或者商品用户很可能早已从其他渠道得到，因此用户不会认为这样的推荐是有价值的。为了弥补基于预测准确度的评价指标的不足，最近相关学者提出了衡量推荐多样性和新颖性的指标[50,55]。

在推荐系统中，多样性体现在以下两个层次，用户间的多样性(inter-user diversity)[56]，衡量推荐系统对不同用户推荐不同商品的能力；另一个是用户内的多样性(intra-user diversity)[57]，衡量推荐系统对一个用户推荐商品的多样性。对于用户 u 和 t，可以用汉明距离(hamming distance)来衡量这两个用户推荐列表的不同程度，具体定义为：

$$H_{ut}(L)=1-\frac{Q_{ut}(L)}{L} \tag{24}$$

式中，$Q_{ut}(L)$ 表示用户 u 和 t 推荐列表中相同商品的个数。如果两个推荐列表是完全一致的，那么 $H_{ut}(L)=0$，反之如果两个推荐列表没有任何重叠的商品则 $H_{ut}(L)=1$。所有的用户对的汉明距离的平均值即是整个系统的汉明距离 $H(L)$。汉明距离越大，表示推荐的多样性越高。

将系统为用户 u 推荐的商品集合记为：$O_R^u=\{\alpha,\beta,\cdots\}$，那么用户 u 的Intra-user diversity定义为：

$$I_u(L)=\frac{1}{L(L-1)}\sum_{\alpha\neq\beta}s(\alpha,\beta) \tag{25}$$

式中，$s(\alpha,\beta)$ 表示商品 α 和 β 的相似度，系统的Intra-user diversity即是所有用户的平均值，其中对于用于 u 来说，商品 α 对该用户推荐结果多样性的贡献为：

$$I_u(\alpha,L)=\frac{1}{L}\sum_{\{\beta\in O_R^u\}\cap\{\alpha\neq\beta\}}s(\alpha,\beta) \tag{26}$$

显然，I_u 越小，表明系统为用户推荐的商品的多样性越高，系统的多样性也就越大。

除了多样性以外，新颖性也是影响用户体验的重要指标之一。它指的是向用户推荐非热门非流行商品的能力。前面已经提到推荐流行的商品纵然可可能在一定程度上提高了推荐准确率但是却使得用户体验的满意度降低了。度量推荐新颖性最简单的方法是利用推荐商品的平均度[58]。推荐列表中商品的平均度越小，对于用户来说，其新颖性就越高。由此得到推荐新颖性指标：

$$N(L)=\frac{1}{ML}\sum_u\sum_{\alpha\in o_R^u}k_\alpha \tag{27}$$

式中，k_α 是商品 α 的度；流行度越低表示推荐的结果越新颖。自信息(self-information)[59]也是可以用来衡量推荐出奇意外程度的指标。对于商品 α，一个随机选取的用户选到它的概率是 $\frac{k_\alpha L}{ML}=\frac{k_\alpha}{M}$，所以该商品的自信息量可以表示为：

$$U_\alpha=\log_2\frac{M}{k_\alpha} \tag{28}$$

系统的自信息量 $U(L)$ 也是所有用户的推荐列表中商品的自信息量的均值。

另外，文献[64]提出了衡量推荐新颖性的指标(unexpectedness，UE)，它的主要思路也很简单。UE认为容易被预测出来的商品对用户来说新颖性较差，而那些不容易被预测出来的商品对用户来说比较新颖。作者用一些比较简单的粗糙的推荐系统(primitive prediction method，PPM)来衡量商品是否容易被预测出来，认为PPM就能预测到的商品的新颖度是比较低的，相反PPM未能预测到的商品对用户来说是比较新颖的。基于以上讨论，用户 u 的UE指标定义为：

$$\mathrm{UE}_u=\frac{1}{N}\sum_{\alpha=1}^{N}\max(\mathrm{Pr}_{u\alpha}-\mathrm{Prim}_{u\alpha},0)\mathrm{rel}_{u\alpha} \tag{29}$$

式中，$\mathrm{Pr}_{u\alpha}$ 表示推荐系统预测用户 u 喜欢商品 α 的概率；$\mathrm{Prim}_{u\alpha}$ 表示PPM系统预测的用户 u 喜欢商品 α 的概率；当且仅当PPM预测出的用户喜欢概率低于待评价推荐系统预测的概率时，商品 α 才被认为是新颖的。其中 $\mathrm{rel}_{u\alpha}\in\{0,1\}$，$\mathrm{rel}_{u\alpha}=1$ 时表示商品 α 确实是用户 u 所喜欢的，相反 $\mathrm{rel}_{u\alpha}=0$ 则表示商品 α 并不是用户 u 所喜欢的。值得注意的是，UE指标并没有考虑商品的推荐排序，也就是说排在第一位的新颖商品和排在第100位的新颖商品对于该指标的贡献都是一样的，这显然是不太合理的，因为用户往往

更关心的是排在推荐列表前面的商品。基于此，作者对UE指标稍做改进提出了UER指标[64]，定义为：

$$\mathrm{UER}_u = \frac{1}{N}\sum_{\alpha}\max(\mathrm{Pr}_{u\alpha} - \mathrm{Prim}_{u\alpha}, 0)\frac{\mathrm{rel}_{u\alpha}}{l_{u\alpha}} \quad (30)$$

式中，$l_{u\alpha}$ 表示商品 α 在用户u推荐列表中的排序值。显然在上式中，新颖的商品排得越靠前，其对系统新颖性的贡献也越大。

以上所述的多样性和新颖性指标虽然计算简单，但是大都比较粗糙有一定的局限性，除了UER考虑了推荐商品的预测排序值外，其他指标均未涉及商品预测排序值以及排序偏差等因素。文献[66-68]就这一问题进行了深入的分析提出了更全面的多样性和新颖性评价框架。除了考虑商品间的相似性外，还考虑了Discovery (用户是否知道该商品)、Relevance (推荐系统预测的用户是否喜欢该商品)、Choice (用户是否喜欢该商品)等3个方面，并由此提出了更广义的多样性指标，其定义为：

$$\mathrm{DRC}_u = \frac{1}{\sum_{\alpha}\mathrm{dc}(l_{u\alpha})}\sum_{\alpha}\mathrm{dc}(l_{u\alpha})\mathrm{Pr}_{u\alpha} f_u(\alpha) \quad (31)$$

式中，$\mathrm{dc}(l_{u\alpha})$ 表示推荐列表排序偏差的折扣函数；$l_{u\alpha}$ 表示商品 α 在推荐列表中的排序；$\mathrm{Pr}_{u\alpha}$ 表示推荐系统预测的用户u喜欢商品 α 的概率；$f_u(\alpha)$ 则可以是普适的衡量对于用户u来说商品 α 多样性或者新颖性的函数。当令 $\mathrm{dc}(l_{u\alpha})=1$ 时相当于不考虑推荐列表排序偏差的影响。当令 $\mathrm{Pr}_{u\alpha}=1$ 时相当于不考虑商品预测排序值的影响，也即是将推荐列表中的所有商品同等看待。在此框架下，我们可以对前面提到的Intra-user Similarity做如下改进：

$$I_u = \frac{1}{\sum_{\alpha}\mathrm{dc}(l_{u\alpha})}\sum_{\alpha}\mathrm{dc}(l_{u\alpha})\mathrm{Pr}_{u\alpha} I_u(\alpha, L) \quad (32)$$

式中，$I_u(\alpha, L)$ 为对于用户u来说商品 α 对其Intra-user similarity的贡献，如式(26)所示。当 $\mathrm{dc}(l_{u\alpha})=1$ 且 $\mathrm{Pr}_{u\alpha}=1$ 时，该指标就退化为原始的Intra-user similarity，即式(25)。显然，这一框架提高了以往多样性和新颖性指标的客观性以及合理性。相对于准确率指标，目前多样性和新颖性指标还不够成熟，但是相信随着推荐系统在商业领域的广泛应用，这方面的研究会不断地发展和完善。

为了给出更直观的比较，将本文所涉及的评价指标总结如表2所示。

表2　推荐系统评价指标简表

评价指标		名称	符号	偏好	是否依赖于推荐列表长度	备　注
准确度	预测评分准确度	平均绝对误差	MAE	小	否	适用于比较关注精确的预测评分的系统
		平均平方误差	MSE			
		均方根误差	RMSE			
		标准平均绝对误差	NMAE			
	预测评分关联	Pearson关联	PCC	大	否	适用于不关注精确预测评分的系统，其中NDMP适用于弱排序
		Spearman关联	ρ	大		
		Kendall's Tau	τ	大		
		基于距离的标准指标	NDMP	小		
	分类准确度	准确率	$P(L)$	大	是	除AUC外，其他不适用于没有明确二分喜好的系统
		召回率	$R(L)$		是	
		准确率提高率	$e_{\mathrm{P}}(L)$		是	
		召回率提高率	$e_{\mathrm{R}}(L)$		是	
		F1指标	$F_1(L)$		是	
		ROC曲线面积	AUC		否	
	排序准确度	平均排序分	RS	小	否	适用于对推荐排序要求严格的系统
基于排序加权的指标		半衰期效用指标	HL(L)	大	是	考虑了具体的推荐排序值，更合理些
		折扣累计利润	DCG(b,L)			
		排序偏差准确率	RBP(p,L)			
覆盖率		预测覆盖率	COV_P	大	否	这些指标单独使用没有意义，应与准确度指标一起考虑。欲计算种类覆盖率指标需要先对商品种类分类
		推荐覆盖率	$\mathrm{COV}_{\mathrm{R}}(L)$		是	
		种类覆盖率	COV_C		是	
多样性		Inter-user diversity	$H(L)$	大	是	
		Intra-user diversity	$I(L)$	小		
新颖性		推荐商品平均度	$N(L)$	小	是	
		系统的自信息量	$U(L)$	大		
		推荐的新颖率	UE	大		
		考虑排序的推荐新颖率	UER	大		

6 总结与展望

本文对现有的推荐系统评价指标进行了系统的回顾，总结了推荐系统评价指标的最新研究进展，从准确度、多样性、新颖性及覆盖率等方面进行多角度阐述，并对各自的优缺点以及适用环境进行了深入的分析。特别讨论了基于排序加权的指标，强调了推荐列表中商品排序对推荐评价的影响。

到目前为止，如何客观、有效的评价推荐系统仍然是一个没有定论的问题。在众多的评价指标中，如何进行选择实际上是非常困难的。推荐系统在某些指标上表现好，在某些指标上表现差，因此很难综合的判断这个系统的好坏。在测试中，人们往往根据推荐系统的具体任务进行指标选择。如何设计一种评价指标能够综合的评价推荐系统的表现是一个巨大的挑战。另外，在离线测试中表现好的系统，在进行在线测试的时候由于受到更多不定因素，例如用户界面环境，用户情绪等的影响，很可能表现很差。尽管做到完全客观全面的评价是非常困难的，但是有一点是肯定的，那就是一个好的推荐系统一定是以用户体验为中心的。用户的体验和反馈是评价推荐系统最真实，最客观，最重要的指标。但是，如何将用户的体验感进行量化也是一个具有挑战性的课题。设计以用户体验为中心的推荐系统，除了提高推荐精确度、多样性、新颖性以及覆盖率等指标以外，还应该考虑以下几方面：

1) 提高推荐算法效率，加强增量算法研究。在信息爆炸的今天，推荐算法的时间效率和空间效率显得越来越重要[60]。面对分分钟产生的海量数据，算法是否能够做出及时的处理。另外算法是否能够抓住用户的即时兴趣，真正的实现时时推荐。在这方面，增量算法的研究至关重要。

2) 提高推荐算法的鲁棒性，减少推荐的干扰因素，净化系统环境。推荐算法的核心就是通过分析用户的历史行为挖掘用户潜在的兴趣，从而为其进行推荐。因此进行准确推荐的一个最基本的前提就是数据是可信的。但是有些恶意用户为了有针对性的提高某个商品的评分而进行一些蓄意的操作。这些行为无疑将破坏推荐系统的正常运转。如何有效的识别出这些恶意用户以及他们制造的垃圾信息对于推荐系统的长期健康发展至关重要。

3) 提高推荐的可解释性，增强用户对推荐系统的信心。有研究发现大多数用户会更倾向于接受可解释性高的推荐[63]。如在为某用户推荐一本书的时候，如果告诉他因为其好朋友也读了这本书，那么该用户接受这个推荐的可能性就更大。

4) 加强推荐系统用户界面的友好程度。有研究表明用户对推荐系统的满意程度还会受推荐系统界面设计的影响[61-62]。一个明亮、整洁、轻松的界面肯定比一个阴郁、杂乱、压抑的界面更能让用户心情愉悦。但是目前还没有考虑此因素的评价指标。

5) 增强用户与系统的互动，从而更加深入细致的挖掘用户的兴趣。在有些参数依赖的推荐算法中[13,53,57]，对于整个系统的最优参数不一定是对每个单独的用户都是最优的。然而，系统针对每个用户学习最优参数是几乎不可能的。一种更可行的方式就是加强用户与系统的互动，让用户通过自身的体验寻找适合自己的参数。如可以在推荐界面上添加一个温度调节阀，高温对应于推荐热门商品，低温对应于推荐冷门商品。已有研究表明，老用户倾向于选择冷门的商品，而新用户更喜欢选择热门商品[65,69]。

综上所述，无论是提高推荐算法效率、鲁棒性或可解释性，加强用户界面的友好程度，还是加强用户与系统的互动，其最终目的都是为了提高用户体验感和满意度。为什么有时候人们更愿意去一些环境优雅的小书屋阅读或购买书籍，而不是选择品种更加齐全的大书城，原因就在于特色的书吧相比千篇一律的书城无论在购物环境还是服务上都能够使顾客更满意。真正的利润不是来自于商品本身，而是来自于用户的购买或点击行为。用户良好的体验感正是促成这种行为的原动力。如何设计以用户体验为中心的推荐系统将成为下一代信息过滤技术的一个核心问题，而相应的系统评价指标的设计也是任重而道远。

参 考 文 献

[1] BAWDEN D, HOLTHAM C, COURTNEY N. Perspectives on information overload[J]. Aslib Proceedings, 1999, 51(8): 249-255.

[2] HWANG M I, LIN J W. Information dimension, information overload and decision quality[J]. Journal of Information Science, 1999, 25(3): 213-218.

[3] EDMUNDS A, MORRIS A. The problem of information overload in business organisations: a review of the literature[J]. International Journal of Information management, 2000, 20(1): 17-28.

[4] EPPLER M J, MENGIS J. The concept of information overload: a review of literature from organization science, accounting, marketing, MIS, and related disciplines[J]. The Information Society, 2004, 20(5): 325-344.

[5] LEE B K, Lee W N. The effect of information overload on consumer choice quality in an on-line environment[J]. Psychology and Marketing, 2004, 21(3): 159-183.
[6] 苏萌, 柏林森, 周涛. 个性化商业的未来[M]. 北京: 机械工业出版社, 2012.
SU Meng, BO Lin-sen, ZHOU Tao. Personalization: the future of business[M]. Beijing: Mechanic Industry Press, 2012.
[7] SCHAFER J B, FRANKOWSKI D, HERLOCKER J, et al. Collaborative filtering recommender system[C]//The Adaptive Web, Lect Notes Comput Sci. Berlin, Heidelberg: Springer-Verlag, 2007, 4321: 291-324.
[8] PAZZANI M J, BILLSUS D, Content-based recommendation systems[C]//The Adaptive Web, Lect. Notes Comput Sci. Berlin, Heidelberg: Springer-Verlag, 2007, 4321: 325-341.
[9] MASLOV S, ZHANG Yi-cheng. Extracting hidden information from knowledge networks[J]. Phys Rev Lett, 2001, 87(24): 248701-248705.
[10] GOLDBERG K, ROEDER T, GUPTA D, et al. Eigentaste: a constant time collaborative filtering algorithm[J]. Information Retrieval, 2001, 4(2): 133-151.
[11] HUANG Z, CHEN H, ZENG D. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering[J]. ACM Transactions on Information System, 2004, 22(1): 116-142.
[12] ZHANG Yi-cheng, MEDO M, REN Jie, et al. Recommendation model based on opinion diffusion[J]. Europhys Letter, 2007, 80(6): 68003-68007.
[13] ZHOU Tao, KUSCSIK Z, LIU Jian-Guo, et al. Solving the apparent diversity-accuracy dilemma of recommender systems[J]. Proceedings of the National Academy of Sciences, 2010, 107(10): 4511-4515.
[14] ZHANG Zi-ke, ZHOU Tao, ZHANG Yi-cheng. Personalized recommendation via integrated diffusion on user-item-tag tripartite graphs[J]. Physica A: Statistical Mechanics and its Applications, 2010, 389(1): 179-186.
[15] BALABANOVIC M, SHOHAM Y. Fab: content-based collaborative recommendation[J]. Comm ACM, 1997, 40(3): 66-72.
[16] SOBOROFF I, NICHOLAS C. Combining content and collaboration in text filtering[C]//Proceedings of the IJCAI'99 Workshop on Machine Learning for Information Filtering. [S.l.]: [s.n.], 1999: 86-91.
[17] TRAN T, COHEN R. Hybrid recommender systems for electronic commerce[C]//Proceedings of Knowledge-Based Electronic Markets, Papers from the AAAI Workshop, Technical Report WS-00-04. Menlo Park: AAAI Press, 2000: 78-83.
[18] GOOD N, SCHAFER J B, KONSTAN J A, et al. Combining collaborative filtering with personal agents for better recommendations[C]//Conference of the American Association of Artificial Intelligence. Menlo Park: AAAI Press, 1999: 439-446.
[19] MELVILLE P, MOONEY R J, NAGARAJAN R. Content-boosted collaborative filtering for improved recommendations[C]//Eighteenth National Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2002: 187-192.
[20] BURKE R. Hybrid recommender systems: Survey and experiments[J]. User Model and User-Adap Interact, 2002, 12(4): 331-370.
[21] YOSHII K, GOTO M, KOMATANI K, et al. An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model[J]. IEEE Transactions on Audio Speech and Language Processing, 2008, 16(2): 435-447.
[22] GIRARDI R, MARINHO L B. A domain model of web recommender systems based on usage mining and collaborative filtering[J]. Requirements Engineering, 2007, 12(1): 23-40.
[23] LÜ Lin-yuan, MEDO M, YEUNG C H, ZHANG Yi-cheng, et al. Recommender systems[DB/OL][2012-02-06]. http://arxiv.org/abs/1202.1112.
[24] HERLOCKER J L, KONSTAN J A, TERVEEN L G, et al. Evaluating collaborative filtering recommender systems[J]. ACM Transactions on Information Systems, 2004, 22(1): 5-53.
[25] 刘建国, 周涛, 郭强, 等. 个性化推荐系统评价方法综述[J]. 复杂系统与复杂性科学, 2009, 6(3): 1-10.
LIU Jian-guo, ZHOU Tao, GUO Qiang, et al. Overview of the evaluated algorithms for the personal recommendation systems[J]. Complex Systemard Complex Science, 2009, 6(3): 1-10.
[26] GUNAWARDANA A, SHANI G. A survey of accuracy evaluation metrics of recommendation tasks[J]. Journal of Machine Learning Research, 2009, 10: 2935-2962.
[27] 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 38(5): 651-661.
LÜ Lin-yuan. Link prediction on complex networks[J]. Jounal of university of Electronic Science and Technology of China, 2010, 38(5): 651-661.
[28] 吕琳媛, 陆君安, 张子柯, 等. 复杂网络观察[J]. 复杂系统与复杂性科学, 2010, 7(2-3): 173-186.
LÜ Lin-yuan, LU Jun-an, ZHANG Zi-ke, et al. Looking into complex network[J]. Complex Systems and Complex Science, 2010, 7(2-3): 173-186.
[29] LÜ Lin-yuan, ZHOU Tao. Link prediction in complex networks: a survey[J]. Physica A, 2011, 390(1150): 1150-1170.
[30] SHARDANAND U, MAES P. Social information filtering: algorithms for automating "word of mouth"[C]// Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems. New York: ACM Press, 1995: 210-217.
[31] BREESE J S, HECKERMAN D, KADIE C. Empirical analysis of predictive algorithms for Collaborative filtering[C]//Proceedings of the 14th conference on Uncertainty in Artificial Intelligence. [S.l.]: [s.n.], 1998, 461(8): 43-52.
[32] HERLOCKER J L, KONSTAN J A, BORCHERS A, et a1. An algorithmic framework for performing collaborative filtering[C]//Proceedings of the 22nd International Conference on Research and Development in Information

Retrieval (SIGIR'99). New York: ACM Press, 1999: 230-237.
[33] RODGERS J L, NICEWANDER W A. Thirteen ways to look at the correlation coefficient[J]. The American Statistician, 1988, 42(1): 59-66.
[34] SPEARMAN C. The proof and measurement of association between two things[J]. The American Journal of Psychology, 1987, 100 (3/4): 441-471.
[35] KENDALL M G. A new measure of rank correlation[J]. Biometrika, 1938, 30(1/2): 81-89.
[36] YAO Y Y. Measuring retrieval effectiveness based on user preference of documents[J]. Journal of the American Society for Information Science, 1995, 46(2): 133-145.
[37] BREESE J S, HECKERMAN D, KADIE C. Empirical analysis of predictive algorithms for collaborative filtering[C]//Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98). San Francisco: Morgan Kaufmann Publishers Inc, 1998: 43-52.
[38] BILLSUS D, PAZZANI M J. Learning collaborative information filters[C]//Proceedings of the 15th National Conference on Artificial Intelligence (AAAI1998). San Francisco: AAAI Press, 1998: 46-54.
[39] BASU C, HIRSH H, COHEN W. Recommendation as classification: using social and content-based information in recommendation[C]//Proceedings of the 15th National Conference on Artificial Intelligence. San Francisco: AAAI Press, 1998: 714-720.
[40] SARWAR B, KARYPIS G, KONSTAN J, et al. Analysis of recommendation algorithms for e-commerce[C]// Proceedings of the 2nd ACM Conference on Electronic Commerce (EC' 00). New York: ACM Press, 2000: 158-167.
[41] SARWAR B M, KARYPIS G, KONSTAN J A, et al. Application of dimensionality reduction in recommender system-a case study[C]//ACM webkDD workshop. New York: ACM Press, 2000, 1(1625): 264-271.
[42] SWETS J A. Information retrieval systems[J]. Science, 1963, 141(3577): 245-250.
[43] CLEVERDON C W, MILLS J, KEAN M. Factors determining the performance of indexing systems[M]. England: Cranfield (Beds) College of Aeronautics, l966.
[44] VAN RIJSBERGEN C J. Information retrieval[M]. MA, USA: Butterworth- Heinemann Newton, 1979.
[45] PAZZANI M J, BILLSUS D. Learning and revising user profiles: the identification of interesting Web sites[J]. Machine Learning, 1997, 27(3): 313-331.
[46] VOORHEES E M, HARMAN D K. TREC: Experiment and evaluation in information retrieval[M]. Cambridge, Mass: MIT Press, 2005: 53-75.
[47] BUCKLEY C, Voorhees E M. Retrieval evaluation with incomplete information[C]//Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2004: 25-32.
[48] MOFFAT A, ZOBEL J, Rank-biased precision for measurement of retrieval effectiveness[J]. ACM Transactions on Information Systems, 2008, 27(1): 2:1-2:27.
[49] HANELY J A, MCNEIL B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve[J]. Radiology, 1982, 143: 29-36.
[50] ZHOU Tao, REN Jie, MEDO M, et al. Bipartite network projection and personal recommendation[J]. Physical Review E, 2007, 76(4): 046115-046121.
[51] SARWAR B, KARYPIS G, KONSTAN J, et al. Item-based collaborative filtering recommendation algorithms[C]// Proceedings of the 10th International Conference on World Wide Web. New York: ACM Press, 2001: 285-295.
[52] JÄRVELIN K, KEKÄLÄINEN J. Cumulated gain-based evaluation of IR techniques[J]. ACM Transactions on Information Systems, 2002, 20(4): 422-446.
[53] LÜ Lin-yuan, LIU Wei-ping. Information filtering via preferential diffusion[J]. Physical Review E, 2011, 83(6): 066119-066130.
[54] CACHEDA F, CARNEIRO V, FERNÁNDEZ D, et al. Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems[J]. ACM Transactions on the Web, 2011, 5(1): 2:1-2:33.
[55] MCNEE S M, RIEDL J, KONSTAN J A. Being accurate is not enough: how accuracy metrics have hurt recommender systems[C]//Proceedings of the CHI 06 Conference on Human Factors in Computing Systems. New York: ACM Press, 2006: 1097-1101.
[56] ZHOU Tao, JIANG Luo-luo, SU Ri-qi, et al. Effect of initial configuration on network-based recommendation[J]. Europhysics Letters, 2008, 81(5): 58004-58007.
[57] ZHOU Tao, SU Ri-qi, LIU Run-ran, et al. Accurate and diverse recommendations via eliminating redundant correlations[J]. New Journal of Physics, 2009, 11: 123008-123026.
[58] ZHANG Zi-ke, LIU Chuang, ZHANG Yi-cheng, et al. Solving the cold-start problem in recommender systems with social tags[J]. Europhysics Letters, 2010, 92(2): 28002-28007.
[59] TRIBUS M. Thermostatics and thermodynamics: An introduction to energy, information and states of matter, with engineering applications[M]. New York: Van Nostrand, Princeton, 1961.
[60] MILLER B N, ALBERT I, LAM S K, et al. MovieLens unplugged: experiences with a recommender systems on four mobile devices[C]//Proceedings of 8th International Conference on Intelligent User Interfaces (IUI'03). Florida: ACM Press, 2003: 263-266.
[61] HELANDER M G, LANDAUER T K, PRABHU P V. Handbook of human-computer interaction[M]. 2ed ed. Amsterdam: North Holland, 1998.
[62] NIELSEN J. Usability engineering[M]. San Francisco: Morgan Kaufmann, 1993.
[63] CHEN Ji-lin, GEYER W, DUGAN C, et al. Make new friends, but keep the old: recommending people on social networking sites[C]//Proceedings of the 27th International Conference on Human Factors in Computing Systems(CHI'09). New York: ACM Press, 2009: 201-210.

[64] MURAKAMI T, MORI K, ORIHARA R. Metrics for evaluating the serendipity of recommendation Lists[C]//Proceedings of the 2007 Conference on New Frontiers in Artificial Intelligence (JSAI'07). Berlin, Heidelberg: Springer-Verlag, 2008: 40-46.

[65] SHANG Ming-sheng, LÜ Lin-yuan, ZHANG Yi-cheng, et al. Empirical analysis of web-based user-object bipartite networks[J]. Europhysics Letters, 2010, 90(4): 48006-48011.

[66] CASTELLS P, VARGAS S, WANG Jun. Novelty and diversity metrics for recommender systems: choice, discovery and relevance[C]//Proceedings of International Workshop on Diversity in Document Retrieval (DDR). Dublin, Ireland: [s.n.], 2011: 29-37.

[67] MOURAO F, FONSECA C, ARAUJO C, et al. The oblivion problem: exploiting forgotten items to improve recommendation diversity[C]//International Workshop on Novelty and Diversity in Recommender Systems. Chicago, Illinios: ACM RecSys, 2011.

[68] VARGAS S. New approaches to diversity and novelty in recommender systems[C]//The Fourth BCS-IRSG Symposium on Future Directions in Information Access. Germany, Koblenz: [s.n.], 2011: 8-13.

[69] ZHANG Cheng-jun, ZENG An. Behavior patterns of online users and the effect on information filtering[J]. Physica A, 2012, 391(4): 1822-1830.

编 辑 蒋 晓