可以划分为三类：

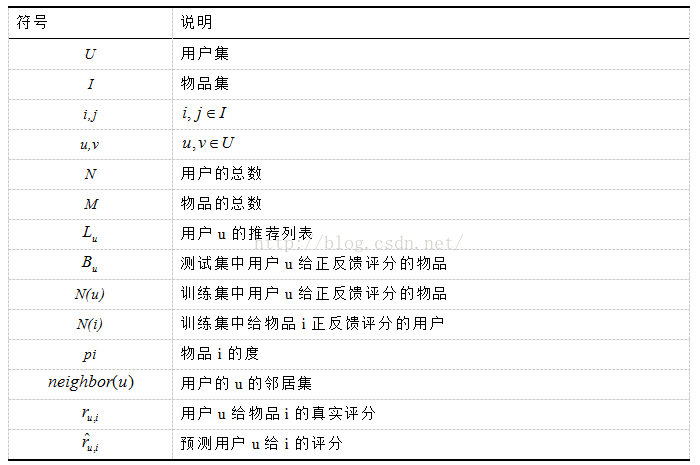
1. 对预测的评分进行评估，适用于评分预测任务。
2. 对预测的item集合进行评估，使用域Top-N推荐任务。
3. 按排名列表对推荐效果加权进行评估，既可以适用于评分预测任务，也适用于

Top-N推荐任务。

这三类度量方法对应的具体评价指标分别为：

* 评分预测指标：如准确度指标：平均绝对误差（MAE）、均方误差根（RMSE）、标准化平方误差（NMAE）；以及覆盖率（Coverage）。
* 集合推荐指标：如精密度（Precision）、召回（Recall）、ROC和AUC。
* 排名推荐指标：如half-life和discounted cumulative gain等。

对推荐系统的研究一个重要的环节是如何评价一个推荐算法的好坏。关于推荐系统评价的研究很多，文献[1,2,3]在不同程度对评价方法进行了总结。评价方法分为离线评估，用户调查，在线评估。由于用户调查和在线评估代价要求高，目前大多数的研究采用的是离线测试。其中文献[3]总结了离线评价中用到的指标，包括准确度指标、基于排序加权的指标、覆盖率、多样性和新颖性等。本文的研究工作采用了几种离线评估指标。将其归纳为准确性指标和非准确性指标。下面分别给予简单介绍。

符号说明：  


1. 准确性指标、

准确性指标是推荐系统中最重要的指标。最常使用的准确性指标包括准确率和召回率。

1. 准确率。推荐列表中用户喜欢的物品所占的比例。针对单个用户u的推荐准确率：



整个系统的准确率为：



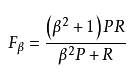
1. 召回率。测试集中有多少用户喜欢的物品出现在推荐列表中。针对单个用户u的推荐召回率：



整个系统的召回率为：



1. F1-Measure。F-Measure又称为F-Score，是IR（信息检索）领域的常用的一个评价标准，计算公式为：



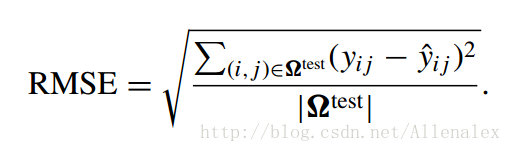
其中β是参数，P是精确率(Precision)，R是召回率(Recall)。

当参数β=1时，就是最常见的F1-Measure了

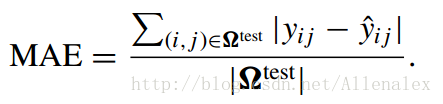


以上的准确性指标是只从整个推荐列表的推荐用户是否喜欢这种二值情况考虑的。除此之外，针对评分（比如0-5分制）的评估主要还有：

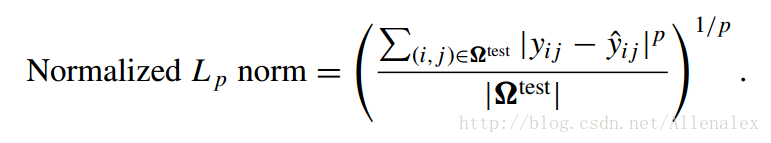
**RMSE（均方误差）**



其中Ωtest为测试集；yij为用户i对物品j的实际评分，y^ij为预测评分。

**MAE（平方绝对误差）：**  


Normalized Lp norm



显然，RMSE,MAE是 Normalized Lp norm的特例。

准确性指标局限性：

不适合评估排序性能。关于排序的性能，推荐大家参考我的另一篇博客： http://blog.csdn.net/allenalex/article/details/78161915

离线的准确率的提高并不意味着实际线上系统效果的提高。

1. 非准确性指标

推荐系统中，除了推荐准确性外，还有其他一些重要的指标。包括推荐的多样性、新颖性、惊喜度和覆盖率等等。本文将这些指标统称为非准确性指标。其中非准确性指标又可以分为用户级非准确性指标和系统级非准确性指标。

* 1. 用户级非准确性指标

1. 个体多样性，用户的推荐列表列内的所有物品的平均相似度：  
   

其中similarity(i,j)可以本博主的另一篇博文《推荐系统中的相似度计算方法总结》介绍的相似度计算方法来计算。求系统中所有用户的推荐列表列内的所有物品的平均相似度的平均值得到整体（推荐列表）列内相似度：

IntraSimilarity值越大，说明用户的推荐列表内的物品之间总体平均相似度越高，也就是系统整体的个体多样性越低。

1. 新颖性[4]。评估新颖性最简单的方法是计算推荐列表中物品的平均流行度：



整个系统的新颖性：



* 1. 系统的非准确性指标。

1. 整体多样性。采用推荐列表间的相似度，也就是用户的推荐列表间的重叠度来定义整体多样性。



1. 覆盖率。覆盖率测量的是推荐系统推荐给所有用户的物品数占总物品数的比例。

