

**中华人民共和国国家计量技术规范**

JJF XXXX—202X

基于用户黏着度的大数据推荐算法

测评方法

Evaluation specification of big data recommendation algorithm

based on customer stickiness

（征求意见稿）

XXXX－XX－XX 发布 XXXX－XX－XX实施

**国家市场监督管理总局发布**

基于用户黏着度的大数据推荐算法测评方法

JJF XXXX-202X

Evaluation specification of big data recommendation algorithm

based on customer stickiness

归口单位：全国数字计量技术委员会

主要起草单位：XXXX

XXXX

参加起草单位：XXXX

XXXX

本规范委托全国数字计量技术委员会负责解释

本规范主要起草人：

XXXX

XXXX

XXXX

参加起草人：

XXXX

XXXX

XXXX

XXXX

目录

[引言 4](#_Toc17113)

[1 范围 5](#_Toc14168)

[2 引用文件 5](#_Toc20984)

[3 术语和定义 5](#_Toc8556)

[3.1 用户黏着度 5](#_Toc32256)

[3.2 大数据 5](#_Toc19623)

[3.3 推荐算法 5](#_Toc9215)

[4 概述 5](#_Toc17091)

[4.1技术指标 5](#_Toc27649)

[4.2 参考数据集 6](#_Toc17365)

[5 技术指标 7](#_Toc21790)

[5.1 算法标识 7](#_Toc23714)

[5.2 用户满意度 7](#_Toc17129)

[5.3 预测准确度 7](#_Toc19660)

[5.4 覆盖率 8](#_Toc8047)

[5.5 置信度 8](#_Toc9044)

[5.6 多样性 9](#_Toc32298)

[6 基于参考数据集的测试流程及结果 10](#_Toc27715)

[6.1 基于参考数据集的测试流程说明 10](#_Toc20375)

[6.2 参考结果 11](#_Toc24692)

[7 测评资料 13](#_Toc29249)

[附录A基于用户黏着度的大数据推荐算法测评报告格式 14](#_Toc20695)

引言

本规范按JJF1071-2010《国家计量校准规范编写规则》要求编制。

本规范结合当前我国大数据推荐算法的技术应用现状，制定了通用的大数据推荐算法技术指标及测评方法。

本规范为首次发布。

基于用户黏着度的大数据推荐算法测评规范

1 范围

本规范规定了基于用户黏着度的大数据推荐相关软件的算法标识、用户满意度、预测准确度、覆盖率、置信度、多样性等技术指标及测评方法。

本规范适用对基于用户黏着度的大数据推荐算法进行测评。

2 引用文件

本规范引用了下列文件：

GB/T 35295 -2017 信息技术 大数据 术语 第2部分：大数据 Big Data

GB/T 8567 计算机软件文档编制规范

GB/T 9385 计算机软件规格需求说明规范

凡是注日期的引用文件，仅注日期的版本适用于本规范；凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本规范。

3 术语和定义

3.1 用户黏着度 customer stickiness

指用户在体验网站、产品或服务后，对其产生的包括但不限于再次使用时长、间隔时长、频率，或产生消费期望的概率。

3.2 大数据 big data

具有体量巨大、来源多样、生成极快、且多变等特征，并且难以用传统数据体系结构有效处理的包含大量数据集的数据。

[来源：GB/T 35295 -2017，2.1.1 ]

3.3 推荐算法 recommendation algorithm

计算机专业的一种算法，通过分析用户的行为、兴趣和需求，利用数学方法和算法预测用户可能再次使用的内容、产品、功能。

4 概述

4.1技术指标

本规范使用“用户满意度、预测准确度、覆盖率、置信度、多样性”五个评价指标对推荐算法的推荐合理性进行测评，评价指标的结果数值越高，说明其算法推荐的越合理。

4.2 参考数据集

本规范依据大数据推荐相关内容将推荐信息的数据集划分成参考数据集Ⅰ和参考数据集Ⅱ，其具体要求说明如下所述。

4.2.1 参考数据集Ⅰ

参考数据集Ⅰ应包含至少数十万条由用户对各类对象（如商品、服务、内容等）给出的评分记录。每条记录应明确包含评分主体（用户）的标识、被评对象的标识以及具体的评分值。评分主体和被评对象应具有自己的属性集合。评分应在明确的数值范围内，如1到5或1到10，以实现标准化。为确保推荐系统的广泛适用性和减少偏见，数据集中的用户属性集合应具有多样性，包括但不限于用户的性别、年龄和职业等信息，以反映不同的用户特征。同时，被评对象的属性集合也应广泛，如所属类别等，以涵盖多样的产品或服务类别。确保每条评分记录都带有时间戳，有助于分析用户偏好的动态变化。所有数据在确保真实性的前提下，必须遵守隐私保护和数据保护法规，对敏感数据进行恰当处理。数据集的文档应详细说明数据的收集、处理方法及数据结构。

例：以MovieLens-1M数据集为例。MovieLens 1M数据集专注于电影推荐，它不仅包含了丰富的用户评分数据，还有电影的元数据和用户的人口统计信息，充分反映了电影推荐领域的特点和需求。该数据集覆盖了多种类型的电影和广泛的用户群体，反映了真实世界中的多样性。这种多样性对于测试推荐算法的泛化能力和处理不同类型用户和项目的能力至关重要。由于是专门为研究目的设计和收集的，MovieLens 1M 数据集在数据的完整性和质量方面具有很高的标准，具有一定的代表性。

4.2.2 参考数据集Ⅱ

参考数据集Ⅱ应包含至少数十万条用户行为数据，如用户的点击流、浏览历史、购买记录等。每条记录必须明确用户主体标识、样本主体标识、对应行为和时间戳。每条记录都配有时间戳，便于分析用户行为随时间的变化。样本主体可以具有自己的属性集合，包括但不限于样本所属类别、所属品牌等信息。所有数据的收集和处理应严格遵守隐私保护法规，确保所有用户信息的匿名性和脱敏处理，以保护个人隐私。为确保推荐系统的代表性和减少偏见，应广泛收集涵盖各种用户群体和行为模式的数据。数据集在保证数据准确性、完整性和一致性的同时，应详细说明数据的收集、处理方法及数据结构。

注：以UserBehavior数据集为例。UserBehavior数据集是阿里巴巴真实电商平台的用户行为数据集，它涵盖了用户的多种行为，包括点击、购买、加入购物车和收藏等，这些复杂的用户行为模式对于研究和模拟现实世界的推荐系统至关重要。该数据集的大规模和高？（多）维度特性使其成为测试和优化大规模推荐系统算法的理想选择，能够处理现实世界推荐系统中常见的挑战，如处理大数据、稀疏性问题和高效计算需求。UserBehavior 数据集的丰富行为数据提供了深入分析用户偏好和行为模式的可能性，具有一定的代表性。

5技术指标

5.1 算法标识

大数据推荐算法应有清晰的软件标识，该软件标识应与大数据推荐算法软件进行关联，并在验证检查时能用于确认大数据推荐算法软件的完整性和唯一性，如果软件有更改，应重新生成软件标识，软件标识及其标识方法应在大数据推荐算法软件测评报告中说明。

5.2 用户满意度

用户满意度是指推荐用户的推荐列表中用户满意样本所占的比例。

对于参考数据集I和II，采用用户满意度 来衡量。具体算法依据公式（1）：

|  |  |
| --- | --- |
|  |  |

式中，是用户对应的推荐列表，是用户在测试数据中具有黏着性的项目集合，其中是测试集中的用户集合。

5.3 预测准确度

预测准确度是指推荐系统预测的样本和真实客户需求相符部分所占的比例。

5.3.1 均方根误差和平均绝对误差

对于参考数据集I，采用均方根误差（误差平方和取均值）和平均绝对误差（误差绝对值和取平均）来衡量。具体算法依据公式（2）和公式（3）：

|  |  |
| --- | --- |
|  |  |
|  |  |

式中，和分别是测试集中的用户集合和样本集合，代表的是测试集中用户对样本的真实评分，而是推荐算法对测试集中用户-样本对生成的预测评分。

5.3.2 命中率

对于参考数据集II，采用命中率来衡量，具体算法依据公式（4）：

|  |  |
| --- | --- |
|  |  |

式中，是测试集中的用户集合，是用户对应的推荐列表，是用户在测试集中具有黏着性的项目集合，是指示函数。

5.4 覆盖率

推荐系统能够推荐出来的用户感兴趣的样本在用户感兴趣的样本集合中的比例。

对于参考数据集I和II，采用覆盖率 来衡量。具体算法依据公式（5）：

|  |  |
| --- | --- |
|  |  |

式中，是测试集中的用户集合，是用户对应的推荐列表,是用户在测试数据中具有黏着性的项目集合。

5.5 置信度

用户对推荐系统预测结果的信任程度。

对于参考数据集I，采用置信度来衡量。不同用户对不同样本有不同的评分。选取一定的置信水平（对于90%的置信水平，对应Z值为1.64；对于95%的置信水平，对应Z值为1.96），对每一个样本都应该存在一个置信区间，表示为在所选置信水平下，该区间包含了真实的平均样本评分。置信区间求解根据公式（6）至（8）。分别计算用户对样本的预测评分在置信区间内和超出置信区间的次数和，使用公式（9）计算置信度。具体算法依据公式（6）至（10）：

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

式中，代表样本 的平均评分，是测试集中的用户集合，表示用户-样本对在测试集中的真实评分，是测试集中对样本存在真实评分的用户数目， 代表样本 评分的标准差，为样本评分的置信区间，是样本集。所有的项目的置信度的平均值即是整个系统的置信度。对于相同的置信水平，值越高，代表信任度越高。

5.6 多样性

推荐系统应尽可能推荐不同的样本。

5.6.1 海明距离

对于参考数据集I，采用海明距离来衡量不同用户之间推荐列表的不同的程度。具体算法依据公式（11）和（12）：

|  |  |
| --- | --- |
|  |  |
|  |  |

式中，表示用户和推荐列表中相同样本的个数，为推荐列表的长度。所有的用户对的海明距离的平均值即是整个系统的海明距离。海明距离越大，表示推荐系统的多样性越高。

5.6.2 香农熵

对于参考数据集II，采用香农熵来衡量用户推荐列表的多样性。具体算法依据公式（13）和（14）：

|  |  |
| --- | --- |
|  |  |
|  |  |

式中，是测试集中的用户集合，是用户对应的推荐列表，是类别的样本在推荐列表中所占的比例。香农熵越高，表明推荐列表覆盖的类别越广泛，多样性越好。

6 基于参考数据集的测试流程及结果

6.1 基于参考数据集的测试流程说明

图6-1描述了基于参考数据集的测试流程结果。

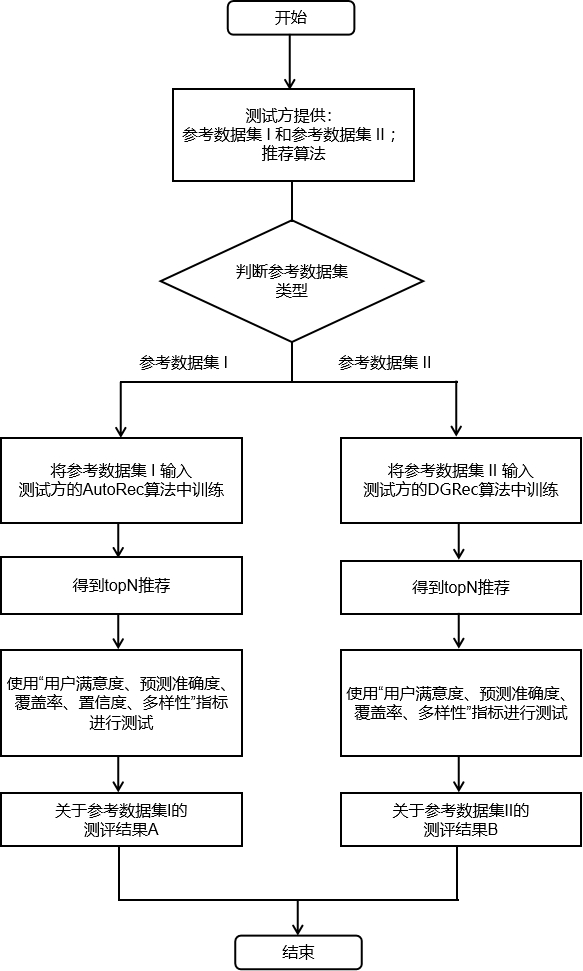


图6-1 基于参考数据集的测试流程图

6.2 参考结果

表6.1显示了基于参考数据集Ⅰ要求的MovieLens-1M数据集在AutoRec算法下的测评结果，评价指标的结果数值越高，说明其算法推荐的越合理。

|  |  |
| --- | --- |
| 评价指标/算法 | AutoRec：top10 |
| 准确度RMSE | 0.8093 |
| 准确度MAE | 0.6484 |
| 满意度CSD | 36.32% |
| 覆盖率Coverage | 22.27% |
| 置信度Confidence | 0.3427 |
| 多样性  Hamming Distance | 0.8567 |

表6.1 MovieLens-1M数据集在AutoRec算法下的参考结果

其中，AutoRec是一种基于自动编码器的推荐系统算法，专门设计来处理用户-项目评分矩阵中的稀疏性问题，并为用户提供个性化推荐。自动编码器是一种无监督的神经网络，它通过学习输入数据的低维表示（即编码）来重构输入数据本身，从而捕捉到数据的潜在结构。在AutoRec中，自动编码器不是直接应用于原始输入数据，而是应用于用户-项目评分矩阵，目标是通过重构评分矩阵来揭示用户偏好和项目属性之间的隐藏关系。AutoRec的核心优点是其能够通过深层神经网络学习复杂的数据表示，从而提高推荐的准确性。此外，由于其基于自动编码器的结构，AutoRec对数据中的噪声和缺失值具有较强的鲁棒性。这种方法不仅能够有效处理稀疏数据集，还能捕捉用户和项目之间复杂的相互依赖关系，为用户提供更加准确和个性化的推荐。

表6.2显示了基于参考数据集Ⅱ要求的UserBehavior数据集在DGRec算法下的参考结果，评价指标的结果数值越高，说明其算法推荐的越合理。

表6.2 UserBehavior数据集在DGRec算法下的参考结果

|  |  |  |
| --- | --- | --- |
| 评价指标/算法 | DGRec：top100 | DGRec：top300 |
| 准确度HR | 0.2966 | 0.4723 |
| 满意度CSD | 0.0042 | 0.0028 |
| 覆盖率Coverage | 4.57% | 9.20% |
| 多样性  Entropy | 0.6409 | 0.6079 |

其中，DGRec算法重点关注增强推荐的多样性，通过优化一个子模函数来寻找一个多样化的邻居子集。从多样化的子集聚集的信息可以帮助我们发现长尾项，并将其反映在聚集表示中，同时添加层注意力模块用来处理过度平滑问题，使得GNN层上的训练更加稳定，并使DGRec能够利用高阶连通性实现多样化。最后，对损失重新加权，降低受欢迎商品/类别的权重，帮助模型更专注于长尾商品/类别。 DGRec 算法将图神经网络技术引入推荐系统，展示了未来推荐系统在建模复杂用户和物品关系方面的潜力和方向。图神经网络的引入为理解和利用用户和物品之间复杂交互提供了新的工具。通过利用图结构，DGRec展示了如何有效地建模和利用用户-用户、物品-物品以及用户-物品之间的复杂关系，这在传统推荐系统中是难以实现的。

7 测评资料

申请单位应提供如下文档资料：

a）算法应符合GB/T 8567以及GB/T9385的相关要求，包括：

1）算法需求规格说明；

2）算法（结构）设计说明；

3）用户手册（应涵盖所有命令集）。

b）可执行程序

1）被测算法可执行程序；

2）加密的可执行程序及加密方法描述；

3）被测评软件安装运行环境、配置参数描述。

c）除上述文档资料之外，制造商还应提交一份关于算法的法律法规符合性、资料完整性和内容一致性的声明。

d）训练数据集规模、数据类型、内容等详细说明文档。

附录A

基于用户黏着度的大数据推荐算法测评报告格式

报告编号：

报告名称：

项目名称：

委托单位：

测评单位：（盖章）

测评日期：

审核日期：

批准日期：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 申请编号 |  | | 算法名称 |  | 软件标识 |  |
| 申请单位名称 |  | | | 申请单位地址 |  |  |
| 联系人 |  | | 电话 |  | 传真 |  |
| 依据的技术文件（代号和名称） |  | | | | | |
| 软件测评的目的和方法 |  | | | | | |
| 参考环境条件及地点 |  | | | | | |
| 样品内容 |  | | | | | |
| 算法开发环境 |  | | | 操作系统 |  | |
| 测评结论 | 推荐合理 | | | | | |
| 测评 |  | | 审核 |  | 日期 |  |
| 提交资料清单 | | | | | | |
| 依据 | 资料分类 | | 资料名称 | | 是否提交 | 备注 |
| 7 a) | 开发文档 | | 1）软件需求规格说明 | |  |  |
| 2）软件（结构）设计说明 | |  |  |
| 3）用户手册 | |  |  |
| 7 b) | 可执行程序 | | 1）被测软件可执行程序 | |  |  |
| 2）加密的可执行程序及加密方法描述 | |  |  |
| 3）被测评软件安装运行环境、配置参数描述 | |  |  |
| 7 c) | 声明 | | 除上述文档资料之外，制造商还应提交一份关于算法的法律法规符合性、资料完整性和内容一致性的声明。 | |  |  |
| 7 d) | 数据集 | | 训练数据集规模、数据类型、内容等简要说明文档。 | |  |  |
| 技术指标 | | | | | | |
| 依据 | | 测评项目 | 标准结果 | | 测试结果  （≥标准结果） | 备注 |
| 5.2 | | 满意度（CSD） | 36.32% | |  | 适用于参考数据集Ⅰ |
| 0.0042  （DGRec：top100） | |  | 适用于参考数据集II |
| 0.0028  （DGRec：top300） | |  | 适用于参考数据集II |
| 5.3 | | 准确度（RMSE） | 0.8093 | |  | 适用于参考数据集Ⅰ |
| 准确度（MAE） | 0.6484 | |  | 适用于参考数据集Ⅰ |
| 准确度(HR) | 0.2966  （DGRec：top100） | |  | 适用于参考数据集II |
| 0.4723  （DGRec：top300） | |  | 适用于参考数据集II |
| 5.4 | | 覆盖率(Coverage) | 22.27% | |  | 适用于参考数据集Ⅰ、II |
| 4.57%  （DGRec：top100） | |  | 适用于参考数据集II |
| 9.20%  （DGRec：top300） | |  | 适用于参考数据集II |
| 5.5 | | 置信度(Confidence) | 0.3427 | |  | 适用于参考数据集Ⅰ |
| 5.6 | | 多样性  (Hamming Distance) | 0.8567 | |  | 适用于参考数据集Ⅰ |
| 多样性  (Entropy) | 0.6409  （DGRec：top300） | |  | 适用于参考数据集II |
| 0.6079  （DGRec：top300） | |  | 适用于参考数据集II |