**基于大数据的用户行为分析及个性化推荐案例**

**——基于物品的电子商务商品推荐**

1. **教学目的与用途**

本案例用于数据挖掘、机器学习与模式识别课程的使用案例。案例依据电子商务网站的大量数据，提取用户行为的特征信息、属性信息等特征，研究用户的兴趣偏好，分析用户的需求和行为，发现用户的兴趣点，通过对用户访问行为数据的分析与处理，采用基于物品的协同过滤算法进行建模分析，最后通过模型评价与结果分析对算法进行改进，从而帮助学生更好地理解协同过滤算法的原理及处理过程以及推荐系统如何帮助用户发现他们感兴趣的物品信息，从而为用户提供个性化的服务。

**2． 涉及知识点**

推荐系统、协同过滤、数据分析、数据清理、数据爬取、数据挖掘。

1. **配套教材**

项亮，推荐系统实战[M]. 人民邮电出版社，2012.

1. **启发思考题**
2. 当偏好矩阵和相似矩阵过大或过于稀疏时有哪些解决方法？
3. 准确度，召回率，F1 分数分别衡量了数据的什么指标？
4. 如何利用Pearson相关系数、夹角余弦相似度法计算相似度？并用代码实现。
5. **分析思路**

案例首先通过讲解关于推荐系统的基本知识，了解推荐算法的主要类型，重点介绍基于协同过滤的推荐算法原理、算法流程以及适用范围，以及Jaccard相似系数，Pearson相关系数和夹角余弦相似度的相关系数计算方法。介绍推荐系统的性能评估指标：精确度、召回率、F1 score。通过代码讲解和实际数据进行代码实现，进一步帮助学生理解算法的原理和流程步骤以及相比于非个性化推荐算法的优缺点。

1. **理论依据及分析**

推荐系统的主要算法可分为两大类:

一、协同过滤算法；二、基于内容数据的过滤算法。基于内容数据的算法适用于冷启动，即用户和项目都比较少的时候，而基于用户行为数据的协同过滤算法在用户和项目较多，数据比较丰富的情况下有较高的准确率。

除此之外，还包括基于社会网络数据的推荐，基于语境（上下文）感知数据的推荐，基于心理学数据的推荐等等。

**6.1协同过滤算法**

协同过滤推荐算法的基本假设是:为用户推荐感兴趣的内容可通过找到与该用户偏好相似的其他用户，将他们感兴趣的内容推荐给该用户。算法可以概括为两步:①采用相似度计算方法计算用户/项目之间的相似度，构造相似度矩阵;②采用相应的算法估计评分，并据此为用户进行推荐。该类方法可为目标用户估计对某一特定项目的评分，也可产生一个推荐列表。进一步可分为基于基于用户的协同过滤算法（user-based CF）和基于内容（项目、物品）的协同过滤（item-based CF）。

**6.1.1 基于用户的协同过滤算法（user-based CF）**

基于用户(user-based)的推荐算法为目标用户$U\_{i}$(i=1，2，…，n)估计给定项目$I\_{j}$ (j=1，2，…，m)的评分$ p\_{i,j}$。该方法首先计算用户$U\_{i}$和其他用户间的相似度，选取其他用户中为$I\_{j}$评过分的用户构成集合$U\_{i}\*$，根据所有$U\_{k}\in U\_{i}\*$对项目$I\_{j}$的评分来估计$U\_{i}$对$I\_{j}$的评分。一个用户喜欢和他具有相似喜好的用户喜欢的项目，两个用户喜欢的项目交集越大，这两个用户越相似。基于用户的推荐方法面对用户数量不断变化的系统，需要经常重新计算用户间的相似度矩阵，因而时间复杂性高，可扩展性较差。

**6.1.2 基于内容的协同过滤（item-based CF）**

基于内容的推荐算法为目标项目$I\_{i}$ (i=1，2，…，n)估计给定用户$U\_{j}$ (j=1，2，…，m)的评分$ p\_{i,j}$。该方法首先计算项目$I\_{i}$和其他项目间的相似度，选取其他项目中为$U\_{j}$评过分的项目构成集合$U\_{i}\*$，根据所有$U\_{k}\in U\_{i}\*$对用户$U\_{j}$的评分来估计$I\_{i}$对$U\_{j}$的评分。它基于这样的假设：一个用户会喜欢与他之前喜欢的内容相似的项目。因此，基于内容的协同过滤推荐关键在于计算物品之间的相似度。

协同过滤推荐算法需要维护一个用户相似度矩阵或内容相似度矩阵，因此对于内容的数目更新速度远远小于用户数目的增长速度的情况，宜采用基于内容的推荐算法。基于内容的算法一般在性能上要优于基于用户的算法。

**6.1.3相似度计算方法**

两个用户兴趣相似度的计算可以有多种方法，常见有Jaccard相似系数, Pearson相关系数和夹角余弦相似度计算。

1. **Jaccard相似度**

$$J\left(A\_{1},A\_{M}\right)=\frac{\left|A\_{1}\bigcap\_{}^{}A\_{M}\right|}{\left|A\_{1}\bigcup\_{}^{}A\_{M}\right|}$$

分母$A\_{1}\bigcup\_{}^{}A\_{M}$表示喜欢的物品1（用户）与物品（用户）M的用户总数，分子$A\_{1}\bigcap\_{}^{}A\_{M}$表示同时喜欢物品（用户）1与物品（用户）M的用户数。

1. **Pearson相关系数**

$$sim\left(x,y\right)=\frac{\sum\_{i\in I\_{x,y} }^{}( r\_{x,i}-\overbar{r\_{x}}) ( r\_{y,i}-\overbar{r\_{y}})}{\sqrt{\sum\_{i\in I\_{x,y} }^{}\left( r\_{x,i}-\overbar{r\_{x}}\right)^{2} \sum\_{i\in I\_{x,y} }^{}\left( r\_{y,i}-\overbar{r\_{y}}\right)^{2} }}$$

其中 ，$ I\_{x,y}$为用户 x 和 y 共同评过分的项目集合，$ r\_{x,i}$表示用户 x 对项目i 的评分。$\overbar{r\_{x}}$表示用户 x 的平均评分。相关系数取值范围是[-1,1]，当余弦值接近±1，表明两个向量具有较强的相似性。当余弦值为0时表示不相关。

1. **夹角余弦相似度**

$$sim\left(x,y\right)=\frac{\sum\_{i\in I\_{x,y} }^{}r\_{x,i} r\_{y,i}}{\sqrt{\sum\_{i\in I\_{x,y} }^{}r\_{x,i}^{2} \sum\_{i\in I\_{x,y} }^{}r\_{y,i}^{2} }}$$

其中,$ I\_{x,y}$为用户 x 和 y 共同评过分的项目集合,$ r\_{x,i}$表示用户 x 对项目i 的评分。$sim\left(x,y\right)$取值范围是[-1,1]，相关系数的绝对值越大，则表明两者相似性越高。

**6.1.4基于物品的协同过滤算法步骤**

1. 构建偏好二元矩阵P；
2. 用相关性度量方法计算不同物品间的相似度，构建物品相似度矩阵S；
3. 对于用户u，若其偏好矩阵$ P\_{u}$元素之和为零，即没有喜欢的物品（冷启动问题），采用其他推荐方法：随机推荐或热门（Popular）算法;
4. 若其偏好矩阵$ P\_{u}$元素之和不为零，针对其偏好矩阵中感兴趣的物品i，分别从相应相似矩阵$ S\_{i}$中选择相似度最高的K个物品，并按如下公式计算用户u对一个物品j的兴趣：

 $P\_{u,j}=\sum\_{i\in N(u)\bigcap\_{}^{}S(j,M)}^{}w\_{j,i}r\_{u,i}$

其中，$N(u)$为用户喜欢的物品的集合，$S(j,M) $是和物品j最相似的M个物品的集合，$r\_{u,i}$表示用户u对物品i的兴趣（对于偏好矩阵为二元矩阵即$P\_{u,j}$只有0或1的值时，可令$r\_{u,i}=1$）；

1. 选择评估模型，对于预测结果进行评估。

**6.2非个性化推荐**

 作为与协同过滤推荐的对比推荐算法，非个性化算法没有考虑用户的过往行为对于推荐预测结果的影响，因而往往推荐效果不如协同过滤推荐；但遇到冷启动问题时，热门推荐也不失为一种良好的补充。

**6.2.1随机推荐**

随机给用户未查看的物品中推荐多个不同的物品。

**6.2.2热门推荐**

即为某用户从过往点击率最高且该用户未产生过行为（浏览点击查看、评分等）的商品中选出若干个推荐给该用户。

**6.3 评估模型**

推荐系统的评价指标有预测准确度、覆盖率、召回率、F1分数（score）、受试者工作特征曲线（ROC）等。其中准确度是最重要也是最常用的推荐系统评测指标，用来度量推荐系统预测的能力。对于top-M推荐，用准确率（precision）召回率（recall）和F1分数来评测推荐系统。

根据不同的推荐模型应当选择不同的模型评测指标，常用的评测指标模型见下表1：

 **表1 评测指标表**

|  |  |  |  |
| --- | --- | --- | --- |
| 数据表现方式 | 模型1 | 模型2 | 模型3 |
| 预测准确度 | $RMSE= \sqrt{\frac{1}{N}\sum\_{}^{}\left( r\_{u,i}-\overbar{r\_{u}}\right)^{2}}$  | $$MAE= \frac{1}{N}\sum\_{}^{}\left| r\_{u,i}-\overbar{r\_{u}}\right|$$ | $precision= \frac{TP}{TP+FP}$  |
| 分类准确度 | $precision= \frac{TP}{TP+FP}$  | $recall= \frac{TP}{TP+FN}$  | $F1= \frac{TP}{TP+FN}$  |

由于用户行为存在不同，例如加入购物车、是否购买、是否评论、评论的内容、点赞等等不同，衡量表示所有行为喜好程度的评分机制也会有所不同。在某些网站的物品需要对其进行评分，因而需要使用的预测准确度评分方式：均方根误差（RMSE），平均绝对误差（MAE）。本例中的数据为用户浏览过的商品，行为可划分为二元选择（浏览过标记为1，未浏览过标记为0），因此可以选用模型3进行衡量。

对于数据的类别可统一分为两类：正类和负类。数据有两种状态：测试集数据和预测结果数据。对一批测试数据进行预测，结果可以分成四种。

TP(True Positive): 原本是正类，预测结果为正类。(正确预测为正类)。FP(False Positive): 原本是负类，预测结果为正类。(错误预测为正类)。TN(True Negative): 原本是负类，预测结果为负类。(正确预测为负类)。FN(False Negative): 原本是正类，预测结果为负类。(错误预测为负类)。

6.3.1准确率

$$precision= \frac{TP}{TP+FP}$$

在预测结果为正类的数据中，有多少数据被正确预测(原本就是正类)。

6.3.2召回率

$$recall= \frac{TP}{TP+FN}$$

在测试集中为正类的数据中，有多少数据被正确预测(预测结果是正类)。

6.3.3 F1分数（score）

$$F1= \frac{TP}{TP+FN}$$

　　F1分数等于精确率和召回率的调和平均值，它同时兼顾了分类模型的[准确率](https://baike.baidu.com/item/%E5%87%86%E7%A1%AE%E7%8E%87/5165407)和[召回率](https://baike.baidu.com/item/%E5%8F%AC%E5%9B%9E%E7%8E%87/560642)。F1分数可以看作是模型[准确率](https://baike.baidu.com/item/%E5%87%86%E7%A1%AE%E7%8E%87/5165407)和[召回率](https://baike.baidu.com/item/%E5%8F%AC%E5%9B%9E%E7%8E%87/560642)的一种加权平均，它的最大值是1，最小值是0。

1. **背景信息**

互联网公司在日常运营中生成、累积的用户网络行为数据。一天之中，互联网产生的全部内容可以刻满1.68亿张DVD；发出的邮件有2940亿封之多（相当于美国两年的纸质信件数量）；发出的社区帖子达200万个（相当于《时代》杂志770年的文字量）。在互联网时代发展的前期，为了应对网络数据的增长催生了搜索引擎和网址分类网站，典型代表是谷歌和雅虎。但随着数据量过大现象的越发明显，用户在搜寻查找自己想要的精确信息时变得越来越困难，在电子商务领域这点显得更加突出。因此，信息过载的问题已成为互联网技术中的一个重要难题，在此背景下催生了推荐系统和推荐相关算法的发展。

1. **关键要点**

本案例的关键要点有：数据处理，数据可视化及分析，交叉验证，性能评估，算法实现。

数据获取与数据处理是解决问题的开端和基础，案例分析的前提便是得到可靠有效的数据。数据获取主要是利用电子商务网站的用户购物行为数据。

数据处理是保证后续算法能够实现的关键，主要包括数据的预处理（清洗），去掉无效的属性特征，对于错误显示的数据进行调整，对异常数据进行处理，以及进行特征工程等。

数据分析是选择合适的算法以及进行性能评估的先决条件，并决定了算法工程代码实现的不同方案，正因为本案例用户记录数达百万之多，且用户数远大于物品数才选择了基于物品的协同过滤算法。

性能评估方法包括均方根误差、平均绝对误差、准确度、覆盖率、召回率、F1分数、ROC等。其中准确度是最重要也是最常用的推荐系统评测指标，用来度量推荐系统预测的能力。由于用户行为存在不同，衡量表示所有行为喜好程度的评分机制也会有所不同。本例中的数据为用户浏览过的商品，行为可划分为二元选择（浏览过标记为1，未浏览过标记为0），因此可以选用准确度、召回率、F1分数进行模型性能评估。

交叉验证指将原数据集划分为测试集和训练集，通过多次迭代测试算法性能的过程。本推荐系统使用的是离线数据，交叉验证更是衡量算法能够有效的必备步骤，本案例通过多次划分数据取平均值，保证了算法的有效可靠。

算法实现不仅仅考虑的是理论算法问题，更要考虑实际工程问题即实际算法运行效率问题，数据大小、种类的不同，编写的代码方案也会不同，高效的代码优化对于解决算法实现问题至关重要。

1. **建议课堂计划**

课前计划：提出启发思考题让同学进行该案例教材预习工作和初步思考。

课中计划：简要的课堂前言，明确主题 (5-8mins)

 分组讨论 (30mins)，告知发言要求

小组发言： (每组5-8mins)

引导全班进一步讨论，并进行归纳总结 (20-30mins)

课后计划：让同学完成思考题作为课后作业，并提供学生用户行为数据集，让学生自主进行案例实现和分析。