电子商务网站用户行为分析及服务推荐

- 按地域研究用户访问时间、访问内容和访问次数等分析主题，深入了解用户对访问网站的行为和目的及关心的内容。
- 借助大量的用户的访问记录，发现用户的访问行为习惯，对不同需求的用户进行相关的服务页面的推荐。

分析方法与过程

目标是对用户进行推荐，即以一定方式将用户与物品之间简历联系。为了更好的帮助用户从海量数据中快速发现感兴趣的网页，在目前相对单一的推荐系统上进行补充，采用协同过滤算法进行推荐。

以用户浏览网页的类型进行分类，然后对每个类型中的内容进行推荐。

- 从系统中获取用户访问网站的原始积累
- 对数据进行多维度分析，包括用户访问内容，流失用户分析以及用户分类等分析
- 对数据进行预处理，包含数据去重、数据变换和数据分类等处理过程
- 以用户访问 html 后缀的网页为关键条件，对数据进行处理
- 对比多种推荐算法进行推荐，通过模型评价，得到比较好的智能推荐模型。通过模型对样本数据进行预测，胡德推荐结果

数据抽取

因为本例是以协同过滤算法为主导，其他的推荐算法为辅，而协同过滤算法的特性就是通过历史数据找出相似的用户或者网页。因此，在数据抽取的过程中，尽可能选择大量的数据，这样就能降低推荐结果的随机性，提高推荐结果的准确性，能更好地发掘长尾网页中用户感兴趣的网页。

数据探索分析

对原始数据中的网页类型、点击次数和网页排名等各个维度进行分布分析，获得其内部的规律。

网页类型分析
点击次数分析
网页排名

数据预处理

数据清洗
数据变换
属性规约

模型构建

长尾网页丰富、用户个性化需求强烈以及推荐结果的实时变化，以及结合原始数据的特点：网页数明显小于用户数。采用基于物品的协同过滤推荐系统对用户进行个性化推荐，以其推荐结果作为推荐系统结果的重要部分。因其利用用户的历史行为为用户进行推荐，可以令用户容易信服其推荐结果。

基于用户和基于物品的协同过滤算法的区别：基于用户的协同过滤是用在用户少、物品多的场景，反之基于物品的协同过滤就是用在用户多、物品少的场景。对用户-物品评分矩阵进行转置，就可以将基于用户和基于物品的协同过滤相互转换。

基于物品的协同过滤系统的一般处理过程：分析用户与物品的数据集，通过用户对项目的浏览与否（喜好）找到相似的物品，然后根据用户的历史喜好，推荐相似的项目给目标用户。图 12-9 是基于物品的协同过滤推荐系统图[⊖]，从图中可知用户 A 喜欢物品 A 和物品 C，用户 B 喜欢物品 A、物品 B 和物品 C，用户 C 喜欢物品 A。从这些用户的历史喜好可以分析出物品 A 和物品 C 是比较类似的，喜欢物品 A 的人都喜欢物品 C，基于这个数据可以推断用户 C 很有可能也喜欢物品 C，所以系统会将物品 C 推荐给用户 C。

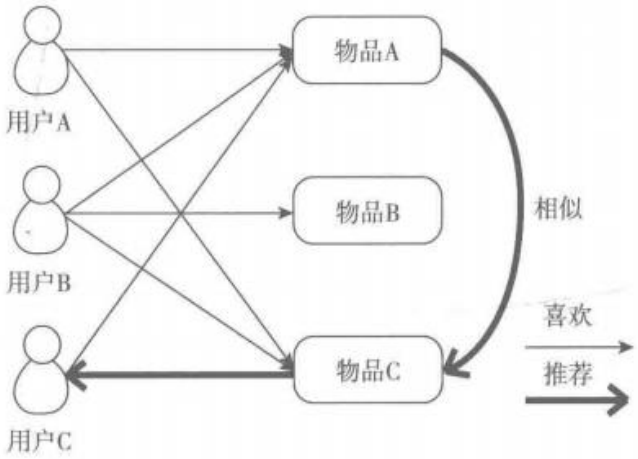

图 12-9　基于物品的推荐系统原理图

根据上述处理过程可知，基于物品的协同过滤算法主要分为两步。

- 计算物品之间的相似度。
- 根据物品的相似度和用户的历史行为给用户生成推荐列表。

方　法	公　式	说　明
夹角余弦	$sim_{1m}=\dfrac{\sum_{k=1}^{n}x_{k1}x_{km}}{\sqrt{\sum_{k=1}^{n}x_{k1}^2}\sqrt{\sum_{k=1}^{n}x_{km}^2}}$ $\left(sim_{1m}=\dfrac{A_1\cdot A_M}{\lvert A_1\rvert\times\lvert A_M\rvert}\right)$	取值范围为 [-1，1]，当余弦值接近 ±1，表明两个向量有较强的相似性。当余弦值为 0 时，表示不相关
杰卡德相似系数	$J(A_1,A_M)=\dfrac{\lvert A_1\cap A_M\rvert}{\lvert A_1\cup A_M\rvert}$	分母 $A_1\cup A_M$ 表示喜欢物品 1 与喜欢物品 M 的用户总数，分子 $A_1\cap A_M$ 表示同时喜欢物品 1 和物品 M 的用户数
相关系数	$sim_{1m}=\dfrac{\sum_{k=1}^{n}(x_{k1}-\overline{A_1})(x_{km}-\overline{A_M})}{\sqrt{\sum_{k=1}^{n}(x_{k1}-\overline{A_1})^2}\sqrt{\sum_{k=1}^{n}(x_{km}-\overline{A_M})^2}}$	相关系数的取值范围是 [-1，1]。相关系数的绝对值越大，则表明两者相关度越高

由于推荐系统是根据物品的相似度以及用户的历史行为对用户的兴趣度进行预测并推荐，因此在评价模型的时候需要用到一些评测指标。为了得到评测指标，一般是将数据集分成两部分：大部分作为模型训练集，小部分数据作为测试集。通过训练集得到的模型，在测试集上进行预测，然后统计出相应的评测指标，通过各个评测指标的值可以知道预测效果的好与坏。

由于推荐系统是根据物品的相似度以及用户的历史行为对用户的兴趣度进行预测并推荐，因此在评价模型的时候需要用到一些评测指标。为了得到评测指标，一般是将数据集分成两部

分：大部分作为模型训练集，小部分数据作为测试集。通过训练集得到的模型，在测试集上进行预测，然后统计出相应的评测指标，通过各个评测指标的值可以知道预测效果的好与坏。

其中，训练集与测试集是通过交叉验证的方法划分后的数据集。通过协同过滤算法的原理可知，在建立推荐系统时，建模的数据量越大，越能消除数据中的随机性，得到的推荐结果对比数据量小要好。但是数据量越大，模型建立以及模型计算耗时就越大。

由于在实际数据中，物品数目过多，建立的用户物品矩阵与物品相似度矩阵是一个很庞大的矩阵。因此，在用户物品矩阵的基础上采用杰卡德相似系数的方法，计算出物品相似度矩阵。通过物品相似矩阵与测试集的用户行为，计算用户的兴趣度，获得推荐结果，进而计算出各种评价指标。

为了对比个性化推荐算法与非个性化推荐算法的好坏，本文选择了两种非个性化算法和一种个性化算法进行建模并对其进行模型评价与分析。两种非个性化算法为：Random 算法和 Popular 算法。其中，Random 算法是每次都随机挑选用户没有产生过行为的物品并推荐给他。Popular 算法是按照物品的流行度，为用户推荐他没有产生过行为的物品中最热门的物品。个性化算法为基于物品的协同过滤算法。利用 3 种算法，采用相同的交叉验证的方法，对数据进行建模分析，获得各个算法的评价指标。

模型评价

离线测试是通过从实际系统中提取数据集，然后采用各种推荐算法对其进行测试，获得各个算法的评测指标。这种实验方法的好处是不需要真实用户参与。

用户调查利用测试的推荐系统调查真实用户，观察并记录他们的行为，并让他们回答一些相关的问题。通过分析用户的行为和他们反馈的意见，判断测试推荐系统的好坏。

顾名思义，在线测试就是直接将系统投入实际应用中，通过不同的评测指标比较不同的推荐算法的结果，比如点击率、跳出率等。

数据表现方式	指 标 1	指 标 2	指 标 3
预测准确度	$RMSE = \sqrt{\frac{1}{N}\sum (r_{ui} - r_{ui})^2}$	$MAE = \frac{1}{N}\sum \lvert r_{ui} - r_{ui} \rvert$	
分类准确度	$precesion = \frac{TP}{TP + FP}$	$recall = \frac{TP}{TP + FN}$	$F1 = \frac{2PR}{P + R}$

结果分析

在协同过滤推荐过程中，两个物品相似是因为它们共同出现在很多用户的兴趣列表中，也可以说是每个用户的兴趣列表都对物品的相似度产生贡献。但是，并不是每个用户的贡献度都相同。通常不活跃的用户要么是新用户，要么是只来过网站一两次的老用户。在实际分析中，一般认为新用户倾向于浏览热门物品，首先他们对网站还不熟悉，只能点击首页的热门物品，而老用户会逐渐开始浏览冷门的物品。因此可以说，活跃用户对物品相似度的贡献应该小于不活跃的用户。所以，在改进相似度的过程中，取用户活跃度对数的倒数作为分子，即本例中相似度的公式为：

$$J(A_1, A_M) = \frac{\sum_{N \in \lvert A_1 \cap A_M \rvert} \frac{1}{\log(1 + A(N))}}{\lvert A_1 \cup A_M \rvert}$$

然而，在实际应用中，为了尽量提高推荐的准确率，还会将基于物品的相似度矩阵按最

大值归一化，其好处不仅仅在于增加推荐的准确度，还可以提高推荐的覆盖率和多样性。由于本例的推荐是针对某一类数据进行，因此不存在类间的多样性，所以本节就不进行讨论。

当然，除了个性化推荐列表，还有另一个重要的推荐应用就是相关推荐列表。有过网购经历的用户都知道，当你在电子商务平台上购买一个商品时，它会在商品信息下面展示相关的商品。一种是包含购买了这个商品的用户也经常购买的其他商品，另一种是包含浏览过这个商品的用户经常购买的其他商品。这两种相关推荐列表的区别为：使用了不同用户行为计算物品的相似性。