《电子商务系统结构》实验指导

实验三：电子商务网站用户行为建模和协同过滤算法

电子商务网站用户行为建模

（1） 基于实验二中选择的电子商务公司，使用其***核心业务***的状态（至少 9 个）用 CBMG 分析其用户行为模型。

（2） 假设我们已经对某电子商务网站的服务器运行日志进行分析，得到客户行为模型的转换概率矩阵 P，如下表所示：

	入口	登录	注册	主页	浏览	退出
入口	0	0.3	0.1	0.5	0.1	0
登录	0	0	0	0.3	0.5	0.2
注册	0	0.4	0	0.3	0.3	0
主页	0	0.2	0.3	0	0.3	0.2
浏览	0	0.2	0.2	0.2	0.2	0.2
退出	0	0	0	0	0	0

请***简单阐述一下解题思路***并求解如下几个问题：

1） 客户行为模型图 CBMG；

2） CBMG 各个状态的平均访问次数；

3） 每次会话的平均会话长度是多少？

（3） 实验报告提交：把 2 个实验一起提交，提交分析简报：文件命名为“学号+姓名”的形式，于 11 月 6 日晚上 12：00 之前发送至 wangmengh@zju.edu.cn

协同过滤算法

推荐系统应用数据分析技术，找出用户最可能喜欢的东西推荐给用户，现在很多电子商务网站都有这个应用。目前用的比较多、比较成熟的推荐算法是**协同过滤**（Collaborative Filtering，简称 **CF**）推荐算法，CF 的基本思想是根据用户之前的喜好以及其他兴趣相近的用户的选择来给用户推荐物品。

A. 什么是协同过滤

协同过滤是利用集体智慧的一个典型方法。要理解什么是协同过滤，首先想一个简单的问题，如果你现在想看个电影，但你不知道具体看哪部，你会怎么做？大部分的人会问问周围的朋友，看看最近有什么好看的电影推荐，而我们一般更倾向于从口味比较类似的朋友那里得到推荐。这就是协同过滤的核心思想。

协同过滤一般是在海量的用户中发掘出一小部分和你品位比较类似的，在协同过滤中，这些用户成为邻居，然后根据他们喜欢的其他东西组织成一个排序的目录作为推荐给你。当然其

中有一个核心的问题：

i. 如何确定一个用户是不是和你有相似的品位？
ii. 如何将邻居们的喜好组织成一个排序的目录？

B. 协同过滤的实现

要实现协同过滤的推荐算法，要进行以下三个步骤：

i. 收集用户偏好
ii. 找到相似的用户或物品
iii. 计算推荐

C. 收集用户偏好

这里的数据指的都是用户的历史行为数据，比如用户的购买历史，关注，收藏行为，或者发表了某些评论，给某个物品打了多少分等等，这些都可以用来作为数据供推荐算法使用，服务于推荐算法。需要特别指出的在于，不同的数据准确性不同，粒度也不同，在使用时需要考虑到噪音所带来的影响。

D. 找到相似的用户或物品

关于相似度的计算，现有的几种基本方法都是基于向量（Vector）的，其实也就是计算两个向量的距离，距离越近相似度越大。在推荐的场景中，在用户 - 物品偏好的二维矩阵中，我们可以将一个用户对所有物品的偏好作为一个向量来计算用户之间的相似度，或者将所有用户对某个物品的偏好作为一个向量来计算物品之间的相似度。下面我们详细介绍几种常用的相似度计算方法：

- 欧几里德距离（Euclidean Distance）

最初用于计算欧几里德空间中两个点的距离，假设 x，y 是 n 维空间的两个点，它们之间的欧几里德距离是：

$$d(x,y)=\sqrt{(\sum(x_i-y_i)^2)}$$

可以看出，当 n=2 时，欧几里德距离就是平面上两个点的距离。

当用欧几里德距离表示相似度，一般采用以下公式进行转换：距离越小，相似度越大

$$sim(x,y)=\frac{1}{1+d(x,y)}$$

- 皮尔逊相关系数（Pearson Correlation Coefficient）

皮尔逊相关系数一般用于计算两个定距变量间联系的紧密程度，它的取值在 [-1，+1] 之间。

$$p(x,y)=\frac{\sum x_iy_i-n\overline{x}\overline{y}}{(n-1)s_xs_y}=\frac{n\sum x_iy_i-\sum x_i\sum y_i}{\sqrt{n\sum x_i^2-(\sum x_i)^2}\sqrt{n\sum y_i^2-(\sum y_i)^2}}$$

s_x, s_y 是 x 和 y 的样品标准偏差。

- Cosine 相似度（Cosine Similarity）

Cosine 相似度被广泛应用于计算文档数据的相似度：

$$T(x,y)=\frac{x\bullet y}{\|x\|^2\times\|y\|^2}=\frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$$

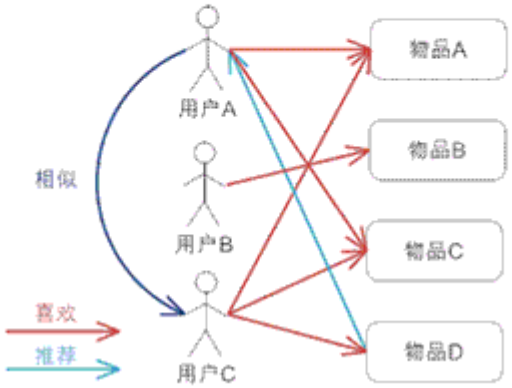
E. 进行推荐

在协同过滤中，有两种主流方法：基于用户的协同过滤，和基于物品的协同过滤。

- 基于用户的协同过滤：

基于用户的 CF 的基本思想相当简单，基于用户对物品的偏好找到相邻邻居用户，然后将邻居用户喜欢的推荐给当前用户。计算上，就是将一个用户对所有物品的偏好作为一个向量来计算用户之间的相似度，找到 K 邻居后，根据邻居的相似度权重以及他们对物品的偏好，预测当前用户没有偏好的未涉及物品，计算得到一个排序的物品列表作为推荐。图 2 给出了一个例子，对于用户 A，根据用户的历史偏好，这里只计算得到一个邻居 - 用户 C，然后将用户 C 喜欢的物品 D 推荐给用户 A。

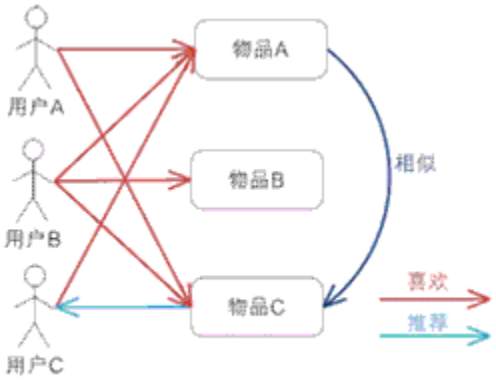
用户/物品	物品A	物品B	物品C	物品D
用户A	√		√	推荐
用户B		√		
用户C	√		√	√



- 基于物品的协同过滤

基于物品的 CF 的原理和基于用户的 CF 类似，只是在计算邻居时采用物品本身，而不是从用户的角度，即基于用户对物品的偏好找到相似的物品，然后根据用户的历史偏好，推荐相似的物品给他。从计算的角度看，就是将所有用户对某个物品的偏好作为一个向量来计算物品之间的相似度，得到物品的相似物品后，根据用户历史的偏好预测当前用户还没有表示偏好的物品，计算得到一个排序的物品列表作为推荐。图 3 给出了一个例子，对于物品 A，根据所有用户的历史偏好，喜欢物品 A 的用户都喜欢物品 C，得出物品 A 和物品 C 比较相似，而用户 C 喜欢物品 A，那么可以推断出用户 C 可能也喜欢物品 C。

用户/物品	物品A	物品B	物品C
用户A	√		√
用户B	√	√	√
用户C	√		推荐



- 基于用户的协同过滤 vs 基于物品的协同过滤

在选择基于用户相似度还是物品相似度的选择问题上，我们需要考虑的是用户和物品的数量。对于新闻，博客或者微内容的推荐系统来说，物品的数量是海量的，同时也是更新频繁的，而用户相对稳定，因此采用基于用户的协同过滤比较合适。而对于电子商务网站来说，用户的数量往往大大超过物品的种类数量，同时物品的数据相对稳定，因此计算物品的相似度不但计算量较小，同时也不必频繁更新。所以单从复杂度的角度，这两个算法在不同的系统中各有优势，推荐系统的设计者需要根据自己应用的特点选择更加合适的算法。本实验是基于物品的协同过滤实验。

本次实验目的：

通过阅读经典论文了解基于物品的协同过滤推荐算法的原理，通过简单的实验尝试加深对算法的理解。

本次实验内容：

1. 阅读论文《Item-Based Collaborative Filtering Recommendation Algorithms》1-3 节，初步了解协同过滤的原理，并打开附件里的 itemBasedCF.zip 文件，用 java 运行。文件中代码已经实现了基于 Correlation-based Similarity 的物品协同过滤算法，实验数据是人工编造的 toy example。运行 main 函数可以看到，该算法可以基于物品的相似度输出用户对未打分的物品的评分预测，如下所示：

```
Sally to sausage: predictedscore 2.02799113489678
Sally to mutton: predictedscore 0.0
Tommy to pork: predictedscore 2.0
Tommy to chicken: predictedscore 0.0
warning: no such user named Curry in the dataset
```

做完这一步后面的推荐环节就相对简单了，针对本次实验，如果要给某个用户推荐一个新的食物，那就在他没评分的物品中挑出预测分最高的推荐给他。而在电子商务推荐的

实际应用中，我们往往会找出该用户预测分最高的前 20 或者前 50 的物品，显示在网页的推荐栏上面。这里同学们需要尝试运行已写好的代码，并尝试修改不同的测试用户来观察输出的预测分，在实验报告中给出截图并分析。

2. 完成 AdjCosineSimilarityMethod.java 未完成的方法，并补全 ItemBasedModel.java 里 line 55 的代码，并在 main 函数里运行并观察实验结果，在实验报告中给出截图并分析。
3. （选做题）阅读论文全文，使用真实数据集按论文的步骤重复论文实验，实验结果与论文结果对比，在实验报告中给出截图并分析。（做出该实验有额外加分，时间来不及这题选做题可以期末前单独提交。推荐使用 MATLAB 或者 python，效率高）
   实验数据下载地址： http://grouplens.org/datasets/movielens/
4. 实验报告提交：把 2 个实验一起提交，提交分析简报：文件命名为“学号+姓名”的形式，于 11 月 6 日晚上 12：00 之前发送至 wangmengh@zju.edu.cn

Reference：

1， https://en.wikipedia.org/wiki/Collaborative_filtering#Types
2， http://www10.org/cdrom/papers/519/
3， http://www.cnblogs.com/luchen927/archive/2012/02/01/2325360.html
4， http://www.ibm.com/developerworks/cn/web/1103_zhaoct_recommstudy2/index.html.
5， Sarwar, Badrul, et al. "Item-based collaborative filtering recommendation algorithms." Proceedings of the 10th international conference on World Wide Web. ACM, 2001.（论文已经在附件中）