基于用户商品交互行为的聚类研究

赵海臣， 刘仕忠， 甘文杰，王麒，姚杰，肖常学
聚美优品 成都研发中心,成都 610041

摘要：本文提出了一种电商大数据环境下基于用户历史商品交互记录的用户、商品聚类方法，能够有效地基于大规模用户的商品交互记录，对商品按照用户心理期望以及实际行为进行类别聚类划分，以及对用户按照用户的心理偏好进行用户偏好聚类划分，并且提出了一系列评估方法。最后，利用成熟电商公司的 SPARK 大数据环境，对聚类结果进行了一系列测试与验证。
关键词：聚类，用户商品交互，机器学习，大数据
中图分类号:TP301.6 **文献标志码：** A **文章编号：**

User-Item Interactive Behavior Based Cluster Method

Zhao Haichen, Liu Shizhong, Gan Wenjie, Wang Qi, Yao Jie, Xiao Changxue
Chengdu R&D, Jumei International Holding, Chengdu 610041, Sichuan

Abstract: This paper has proposed a cluster method which could cluster users or items only with the large amount of deposited user-item interactive logs in an e-commercial company. It could explore the users' psychological division trends towards the items in the item cluster, or cluster the users by their interests in the user cluster. We also presented a series of assessing methods to assess the clustering results. At last, with "SPARK" big data environment and trillions of logs, we made a series of tests and validations to our method.
Keywords: Cluster, User-Product Interaction, Machine Learning, Big Data

0. 引言

用户、商品聚类在互联网领域有着重要的应用，例如用户、商品标签系统、个性化推荐系统、CRM 客户关系管理等方面[1~4]。大部分现有的用户、商品聚类往往是基于文本、标签、用户 RFM 行为等的客观维度的分析与聚类[5~6]，具有较强的客观性，但没有太考虑用户的主观兴趣方面因素。在商业活动中，用户的心理区分因素却是重要的用户、商品区分因素。例如，类似于著名的啤酒与尿布案例[7]，啤酒与尿布在客观因素上是没有任何联系的商品，但是通过大量的用户行为分析，将啤酒与尿布放在一起了(类似聚类过程)，并从这种基于用户商品主观交互记录的聚类中获得了一定的商业利益。

用户商品的交互行为是用户心理的主观反应，对商品来说，用户每次浏览行为无论有意与否，其潜意识中通常有一定的物品倾向，例如非母婴类用户不会有兴趣去浏览母婴类的商品；如果是口红控的女生，即便是随便浏览，口红的点击率 CTR(Click Through Rate)也要明显高于其它的商品，若是有明确购物目标的一次访问 SESSION，其购物目标以及相似的其它商品(货比三家)会有明显的高 CTR。通过对这些用户的行为进行分析聚类，我们可以还原出一个商品在用户心目中距离最近的其它商品，这个距离不一定是我们客观定义的各种类别，但是它却反应了用户的心理期望。

在完善推荐系统的过程中，我们意识到 CF(Collaborative Filtering)基于邻域的协同过滤[8]中所用到的用户、商品距离计算公式可以进一步用于的用户、商品聚类。在经过一段时间聚类尝试与改进之后，得到了一种具有较好效果的根据用户商品交互记录进行聚类方法。

1. 用户商品交互行为

用户商品交互信息，主要包括浏览记录、订单记录、购物车记录、收藏记录等，统一整理清洗成简洁的四列用户商品交互信息”用户编号，商品编号，兴趣分值，行为日期”，接下来所有的距离计算与聚类计算完全基于这四列数据展开。

在数据清洗过程中，我们需要考虑行为 Session 粒度，即多久时间范围内的事件是相关的。这需要考虑行为聚类假

赵海臣，男，湖南长沙人，机器学习工程师，硕士，主要研究方向为数据挖掘、人工智能。EMAIL: haichenz@jumei.com。

设，我们基于用户行为的商品聚类的假设是，用户一段时间内的浏览购买行为是有焦点兴趣的，即大部分是有明确购买目的，所以一个用户一段时间内的交互商品是相似的，若两个商品在用户的同一个 SESSION 内被同时访问，则表示两个商品有较大可能相似，再通过大量用户分析，获得可靠的商品相似度。因此我们定义的商品相似度考察 SESSION 时长按天划分，因为一个用户一天之内的兴趣是通常是较明确的；而用户的聚类假设是，用户是有兴趣偏好的，比如刚生孩子的用户，将会有明确的母婴方面的兴趣爱好，近期一段时间内的交互商品会有明显的母婴倾向分布，在用户的聚类中，SESSION 粒度可定义为较长时长，如一个月。

由于用户、商品都处于千万级别，并且任何一个用户与绝大部分商品都没有交互记录，因此，我们的聚类算法需要能够有效处理这种超高维度的稀疏数据。

2. 交互行为中距离的定义

聚类中最重要的一点是距离的计算，只有定义好了距离计算方式，才能定义簇域。常用的距离计算公式有余弦相似度、皮尔逊相似度、修正余弦相似度等。假设用户是有兴趣偏好的，即其交互的商品是符合用户兴趣的商品：用户 A 与用户 B 有越多共同的交互商品，则表明这两个用户的兴趣偏好越相似，距离越近；同样，商品 A 与商品 B 有越多共同的交互用户，则表明这两个商品越近似，距离越近。

2.1 布尔型余弦相似度

只考虑用户与商品的交互记录布尔值，而不考虑交互次数、时间等其他因素，我们可以使用布尔型余弦相似度：

$$BoolCos_{uv} = \cos A = \frac{\vec{u}\cdot\vec{v}}{\|\vec{u}\|\cdot\|\vec{v}\|} = \frac{|N(u)\cap N(v)|}{\sqrt{|N(u)||N(v)|}} \quad (1)$$

其中 $N(u)$ 、 $N(v)$ 分别代表用户 u、v 的交互商品列表。

相对于标准余弦相似度，由于一个商品对应的用户数量巨大，且用户的浏览次数大部分为区分不大的<10 小值，只考虑用户是否浏览过的 Boolean 值，能够高效地利用大数据量，并且运算速度相对较快。

2.2 标准余弦相似度

我们可以利用用户与商品的交互频次与时间因素获得用户到商品对应的”兴趣评分”，即获得”用户编号，商品编号，兴趣评分”。每个用户因此获得更细节的兴趣区分尺度，我们可以使用基于向量夹角的余弦相似度来计算用户到用户、商品到商品的距离：

$$Cos_{uv} = \cos A = \frac{\vec{u}\cdot\vec{v}}{\|\vec{u}\|\cdot\|\vec{v}\|} = \frac{\sum Rating_{ui} * Rating_{vi}}{\sqrt{\sum Rating_{ui}^2 * \sum Rating_{vi}^2}} \quad (2)$$

其中 $Rating_{ui}$ 、 $Rating_{vi}$ 分别代表用户 u 到商品 i，用户 v 到商品 i 的兴趣评分。

2.3 修正余弦相似度

标准余弦相似度仅考虑向量维度方向上的相似而没考虑到各个维度的量纲的差异性，所以在计算相似度的时候，做了每个维度减去均值的修正操作，即修正余弦相似度[9]：

$$AdjustCos_{uv} = \frac{\sum (Rating_{ui} - Avg(Rating_i)) * (Rating_{vi} - Avg(Rating_i))}{\sqrt{\sum (Rating_{ui} - Avg(Rating_i))^2 * \sum Rating_{vi} - Avg(Rating_i))^2}} \quad (3)$$

2.4 皮尔逊(Pearson)相似度

皮尔逊相关系数主要考虑线性相关性，定义为两个变量之间的协方差和标准差的商，所以自然的考虑的是均值的修正操作[3]。

$$Pearson_{uv} = \frac{\sum (Rating_{ui} - Avg(Rating_u)) * (Rating_{vi} - Avg(Rating_v))}{\sqrt{\sum (Rating_{ui} - Avg(Rating_u))^2 * \sum Rating_{vi} - Avg(Rating_v))^2}} \quad (4)$$

余弦相似度、修正余弦相似度、皮尔逊相似度公式很相似，不同的是，修正余弦相似度与皮尔逊相似度相对于标准余弦相似度多了去中心化的过程，一个是针对维度去中心化，一个是针对变量去中心化。

几种常用计算方式各有利弊，但计算距离的本质都一样，后文以皮尔逊(Pearson)相似度为标准计算距离。

3. 聚类过程

有了距离计算方法后，我们可以利用距离算法来研究下一步的聚类。

常用聚类有三条技术线路[10]，一是类似 K-means 的基于原型、划分的聚类技术线路；二是基于层次凝聚的聚类技术线路；三是基于密度的聚类技术线路。

由于用户数量、商品数量达到千万级别，因此在实际应用中，我们需要考虑运算可行性问题。基本 K-means 的空间需求是适度的，因为只需要存放数据点和质心，存储量为 $O((m+K)n)$，其中 m 是数据点(用户)，n 是属性数(商品)，同时时间复杂度 $O(I\times K\times m\times n)$ 也是适度的，基本

与数据点的个数线性相关；基本层次凝聚技术的空间复杂度为$O(m^2)$，时间复杂度为$O((m^2)\times\log m)$，在大数据环境下层次凝聚技术的成本极其昂贵；而基于密度的聚类线路在这种千万维级别的高维空间中基本无法正常使用。

考虑实际效率以及超高维空间聚类可行性上，我们选择基于原型、划分的技术路线进行聚类研究。与 K-means 聚类类似，我们的聚类方法主要分为三个基础步骤：

1. 生成初始质心；

2. 计算各个点到质心的距离，并更新点的归属簇；

3. 更新质心坐标。

伪代码表示如下：

begin

生成初始质心;

Repeat n:

计算各个点到质心的距离;

更新点的归属簇;

更新质心坐标;

End

以下过程按照用户聚类方式描述。

3.1生成初始质心

我们初始质心生成采用的是随机采样法，k 初始值，将随机选取 k 个点作为初始质点。但是，点在高维空间的分布极其稀疏，大部分用户或商品之间并不会有交集，在这些没有交集的点之间，距离计算失效。

3.2计算点到质心的距离

使用 pearson 相似度计算各个用户 u 到 k 个质心的距离

$$Pearson_{uv}=\frac{\sum(Rating_{ui}-Avg(Rating_u))*(Rating_{vi}-Avg(Rating_v))}{\sqrt{\sum(Rating_{ui}-Avg(Rating_u))^2*\sum Rating_{vi}-Avg(Rating_v))^2}} \quad (5)$$

这样，每个用户(点)u 获得了自己与 k 个质心的距离$PearsonSim(u,k)$，再取每个用户 u 距离最近的一个质心对应的簇作为自己的簇 c。

$$Cluster(u,c)=Nearest(PearsonSim(u,k)) \quad (6)$$

3.3更新质心坐标

更新质心坐标在这种高维度中是有技巧性的，因为一个簇中的用户的商品交互$Rating_{ui}$虽然在大部分商品上会集中，但是各个用户的细微行为差异将覆盖大部分其它的商品维度，导致原本稀疏的质心维度变得不再稀疏(将在大部分维度有个较小但不为 0 的值)，这除了将导致簇与簇之间的干扰外，最主要的，它将严重增加计算复杂度，在大数据环境下是极其敏感的，一般一个用户的交互商品(非 0 的维度)为 100 维左右，一个簇中的用户整体交互商品一般可以覆盖整体商品的 50%以上达到百万级别，导致运算量增加上万倍。因此，为了防止质心维度发散，我们需要同时处理维度的稀疏性。

3.3.1 质心维度坐标值更新

取簇内所有用户 u 在维度 i 的均值作为新的质点维度 i 的 Rating 值

$$Rating_{Centroid-i}=\frac{\sum_{u\in Centroid}Rating_{ui}}{count(u\in Centroid)} \quad (7)$$

3.3.2 维度稀疏性控制

计算簇中用户非零维度数量的平均值$DimenNum_{Centroid}$，再根据 1 中所更新的维度坐标值，保留该簇质心维度值中的最大$DimenNum_{Centroid}$个维度，将其它维度归 0。

$$DimenNum_{Centroid}=\left\lceil\frac{count_{u\in Centroid}(Rating_{ui}\neq 0)}{count(u\in Centroid)}\right\rceil \quad (8)$$

If

$$Rank_{Centroid}(Rating_{Centroid-i})>=DimenNum_{Centroid} \quad (9)$$

Then $Rating_{Centroid-i}=0$ (10)

通过将簇心的维度稀疏度控制在簇平均水平，聚类过程将能够同时在维度级别自动搜索密集维度的能力。另外，通过严格控制了各个簇的维度范围，也降低了维度发散对其它簇聚类过程的干扰。

4. 评估方法

4.1 簇质量指标

通常针对欧几里得空间中的聚类，采用误差的平方和(Sum Of Squared Error, SSE)作为聚类效果评估函数，而对于非欧几里得空间中的聚类，往往使用与 SSE 类似的总凝聚度(total Cohension)作为聚类效果评估函数：

$$TotalCohen\,sion = \sum_{i=1}^{K}\sum_{x\in C_i} Dist(x,c_i) \tag{11}$$

由于 Pearson 相似度的取值区间是[0,1]，与欧几里德距离概念相反，值越大表示距离越近，为了符合我们常理上距离越近值越小的直觉，所以距离取了 1 的差值：

$$Dist(x,c_i) = 1 - PearsonDist(x,c_i) \tag{12}$$

为了使得值更直观，我们采用平均凝聚度作为聚类效果评估函数，使得评估值在[0，1]，越小表示簇质量越高：

$$AvgCohension = \frac{1}{N}\sum_{i=1}^{K}\sum_{x\in C_i}(1 - PearsonDist(x,c_i)) \tag{13}$$

4.2 收敛速度指标

每一次迭代都会计算与上一次迭代的簇质量之间的收敛差异比例，作为收敛速度指标，收敛速度也是我们停止迭代的一个重要参考指标，当收敛速度小于某个阈值，我们即认为簇划分已经趋于一个基本稳定的水平。

$$ConvRate = \frac{AvgCohension_{LastTurn} - AvgCohension_{ThisTurn}}{AvgCohension_{LastTurn}} \tag{14}$$

4.3 点覆盖率

由于维度空间过大，所以必定存在众多未被聚类的孤立点，即与任何簇心都没有任何交集，在稀疏数据计算过程中这些点将自动被清除。所以，聚类点覆盖率也是我们关心的一个指标：

$$SampleCoverage = \frac{N_{ClusteredSamples}}{N_{TotalSamples}} \tag{15}$$

4.4 熵(Entropy)与纯净度(Purity)

基于一些现有的标签，例如对于商品来说，用户对商品的行为往往会有类别与品牌的偏好，这样，对于商品聚类结果，我们可以基于簇的主要聚集特征是类别还是品牌，判断该簇是类别簇还是品牌簇，再基于这些标签计算出簇熵[11]与簇纯净度，用于聚类效果评估。

单独簇的熵：

$$H(\omega) = -\sum_{c\in C} P(\omega_c)\log_2 P(\omega_c) \tag{16}$$

其中 c 代表所有的类别，$P(\omega_c)$ 是对应的类别在簇中的概率，我们可以用比例来代替这个概率。

$$P(\omega_c) = \frac{|\omega_c|}{n_\omega} \tag{17}$$

$$H(\omega) = -\sum_{c\in C}\frac{|\omega_c|}{n_\omega}\log_2\frac{|\omega_c|}{n_\omega} \tag{18}$$

熵总和：

$$H(\Omega) = \sum_{\omega\in\Omega} H(\omega)\frac{N_\omega}{N} \tag{19}$$

其中 N_ω 为该簇中点的个数，N 是所有点的个数。

纯净度：

$$P_\omega = \max(P(\omega_c)) \tag{20}$$

$$purity = \sum_{\omega=1}^{K}\frac{N_\omega}{N}P_\omega \tag{21}$$

5. 实验结果与分析

为了有直观可视的效果评判，我们使用该算法进行商品聚类，因为相比于用户，商品具有很多有利于主观判断的文字信息。

用于实验的计算机资源分配以及软件环境：1TB 内存，300 CPU 核心单元，SPARK 1.6.3，约 100 亿条用户-商品交互记录。

我们使用不同初始点数量，100，500，1000，2000，5000，10000 进行分别实验，记录每次的迭代循环的平均凝聚度、收敛速度，点覆盖率，以及结束后的熵和纯净度，每种数量初始点分别进行 10 个不同随机种子的聚类，所有数据取 10 个不同随机种子结果的平均值。

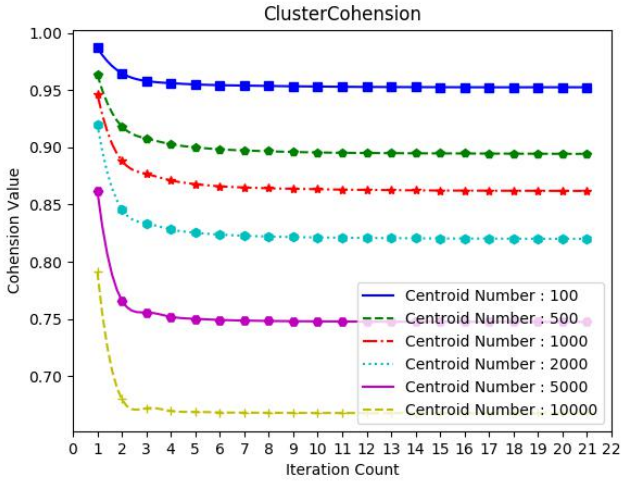

图 1. 不同簇数量下的迭代过程凝聚值

由图 1 可以看到，簇的收敛过程比较迅速，5 次迭代之后就趋于平缓，说明聚类过程非常迅速。而 100、500、1000、2000、5000、10000 的簇数量下收敛稳定后的凝聚度呈现出明显的簇数量越多，凝聚度越大。因此合理簇的数量应该大于 5000。

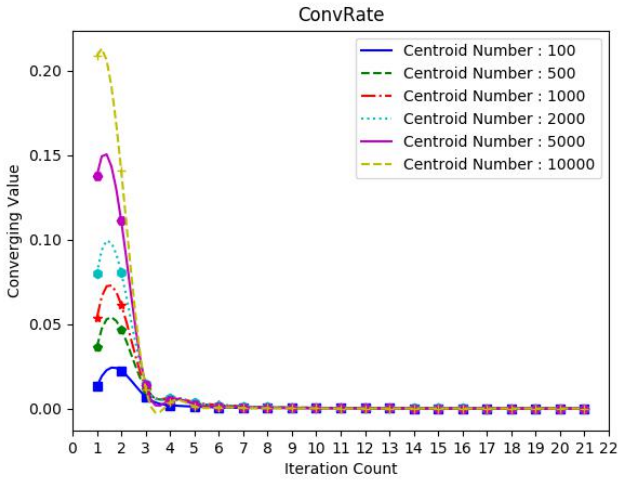

图 2. 不同簇数量下的迭代过程收敛速度

由图 2 可以看到，簇的收敛速度在前 3 轮较快，10 次迭代之后就基本稳定。

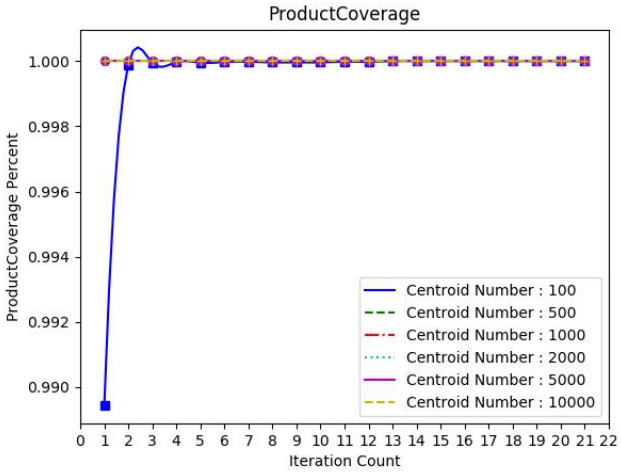

图 3. 不同簇数量下的迭代过程点覆盖率

由图 3 可以看到，簇的覆盖率基本都能覆盖接近全站的所有商品，而 100 个簇的聚类过程在最开始的时候虽然没有覆盖所有商品，但在聚类过程中也覆盖了接近所有的商品。

由图 4 可以看到，簇数量越大，熵值降得越来越平缓。

由图 5 看到，簇数量越大，簇的纯净度越高。由于这些簇的纯净度计算所依赖的标签是按照商品的品牌与分类标签近似计算的，而用户的行为本身是没有客观标签能对应上，因此，实际纯净度应该大于计算出的纯净度。

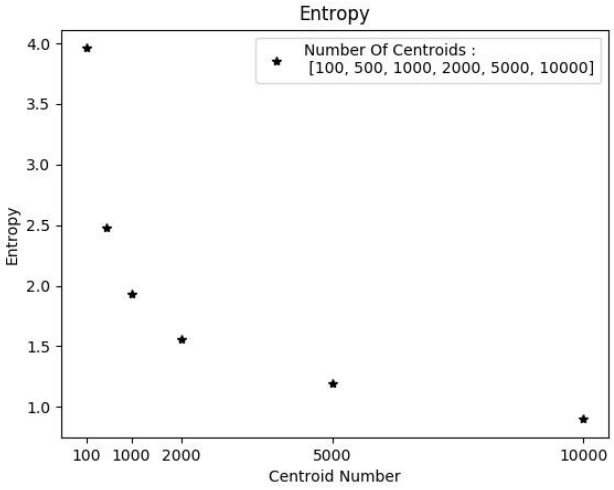

图 4. 不同簇数量下的聚类熵分布

表 1　不同初始簇数量下的熵与纯净度

初始点个数	熵	纯净度
100	3.960585	0.383530
500	2.482221	0.587279
1000	1.928328	0.658170
2000	1.559428	0.702143
5000	1.197476	0.738049
10000	0.899071	0.899071

表 2　商品簇例子

品牌簇例子 —— 贝德玛	类别簇例子 —— 项链
法国•贝德玛 (Bioderma)舒妍多效洁肤液 500ml	Mbox 项链施华洛世奇元素爱情瓶
法国•贝德玛（Bioderma）净妍控油洁肤液 500ml	Mbox 项链 925 银施华洛世奇锆石
法国•贝德玛 (Bioderma)舒妍多效洁肤液 250ml	中国•Mbox 原创韩版潮流小鹿吉祥手链
法国•贝德玛 (Bioderma)净妍控油洁肤液 250ml	中国•今上珠宝 18K 金项链黄金色锁骨链
法国•贝德玛（BIODERMA）控油净妍卸妆水 250ml	中国•Mbox925 银施华洛世奇锆石公主方项
法国•贝德玛（BIODERMA）舒妍卸妆水 250ml	中国•阿梵尼 925 银饰太阳的项链女
法国•贝德玛（Bioderma）舒妍卸妆水 500ml	中国•阿梵尼 925 银太阳吊坠时尚项链女
法国•贝德玛（Bioderma）净妍控油卸妆水 500ml	中国•今上珠宝 925 银项链白金色
法国•贝德玛（BIODERMA）控油净妍卸妆水 100ml	中国•DAZZLING AGE 简约星星钛钢玫瑰金项链
法国•贝德玛（BIODERMA）控油净妍卸妆水两只装（100mlx2）	中国•妍韵珠宝 S925 银 时尚树叶手链
法国•贝德玛（Bioderma）舒妍卸妆水两只装（100mlx2）	中国•DAZZLING AGE925 纯银可爱猫咪珍珠项链
法国•贝德玛 (Bioderma)净妍洁肤液 10ml +专用化妆棉 100 枚（塑袋装）	中国•戴拉 雪花项链 925 银
法国•贝德玛 (Bioderma)舒妍修护爽肤水 250ml	中国•银时代星织梦系列星星项链
法国•贝德玛(Bioderma)净妍控油卸妆水套装(500ml+100ml)	中国•Mbox 项链女 925 银四叶草
法国•贝德玛(Bioderma)控油净妍卸妆水两只装(250ml*2)	中国•Mbox 项链女 925 银钻石奇缘
法国•贝德玛(Bioderma)舒妍卸妆水两只装(250ml*2)	中国•Mbox925 银施华洛世奇锆石圆心项链
法国•贝德玛(Bioderma)舒妍卸妆水套装(500ml+100ml)	rebecca minkoff 瑞贝卡•明可弗水晶镶嵌项链 1142095583
法国•贝德玛（Bioderma）闺蜜分享套装（舒妍卸妆水 250ml+净妍控油卸妆水 250ml)	中国•今上珠宝　925 银太阳项链女锁骨链
。。。	中国•念想爱情双层短款锁骨颈链
	中国•海洋之谜手链
	中国•秘密之约颈链
	。。。

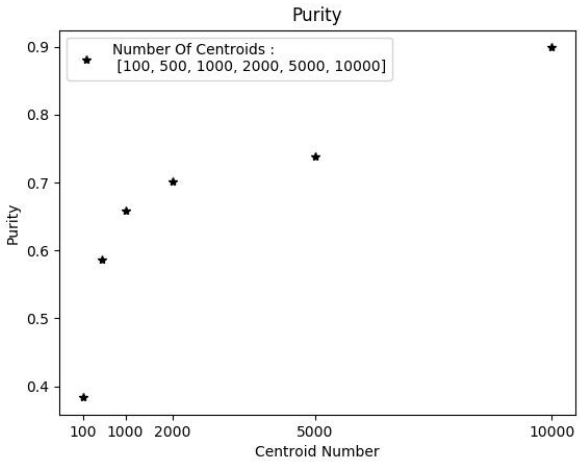

图 5. 不同簇数量下的聚类纯净度分布

当簇 10000 的时候整体聚类纯净度达到了 0.9 以上(簇内相同品牌或者类别的比例达到了 90%以上)，说明 10000 的簇数量已经能够通过用户行为达到细分品牌与类别的级别。

表 2 是聚类的结果随机抽取的典型簇，由于聚类过程完全只使用 uid 与 pid 的交互记录，没有任何语意因素，体现了用户内心对商品的关注边界。因此可以发现，用户对美妆类商品的关注点主要以

1. 品牌，对于知名品牌的商品来说，用户往往有明显的品牌倾向性；

2. 类别，对于普通品牌的商品来说，用户关注点有明显的功能性倾向；

3. 非客观区分，用户的行为意识并不一定能用客观世界的标签对应上，有时候甚至商品促销也可能导致一些商品的聚集。

6. 结论与未来的工作

基于用户商品交互行为可以完成用户、商品的聚类。用户商品的交互行为是用户心理的主观反应，基于用户交互行为对商品的聚类实际上挖掘出了商品的用户心理区分，是一种更主观且贴近消费者的分类体现。

另外，通过对行为聚类发现，用户对商品的关注点主要以品牌或类别作为区分。根据簇的内容可以将聚类结果的簇分为 3 类：

i. 品牌类簇：知名品牌的商品往往是基于品牌的簇，即簇中的商品往往是同一个品牌的；

ii. 类别簇：例如指甲钳、咖啡、绿茶之类的商品，往

往是类别簇，即簇中的商品往往是相同类别的；

iii. 非客观标签簇：基于用户的行为意识聚类并不一定能对应上现实世界的客观标签，一些特殊簇中的联系需要基于心理学与现实进行分析与判断。例如之前提到的啤酒与尿布这样的案例。同样一些商品若进行大量捆绑销售式的促销，也会造成这些非客观关联商品的聚集。

这个结论同时也可以揭示广告的重要性——用户对知名品牌的倾向性相当明显，用户往往具有相当的品牌忠实度，尤其是遍地都是同质化商品的情况下，有个信任的品牌可以避免用户无谓的比较选择。

实现基于用户或者商品的主观兴趣偏好聚类后，我们可以使用这些兴趣聚类结果进行更符合用户兴趣的精准营销，或者对商品的促销、专场进行更合理的规划。由于该版聚类算法较为简洁，实际上留有较大的提升空间。继续优化聚类效果，以及对簇进行自动标签也是我们未来的研究方向。

参考文献

[1] Li F, Wang S, Liu S, et al. SUIT: a supervised user-item based topic model for sentiment analysis[C]// Twenty-Eighth AAAI Conference on Artificial Intelligence. AAAI Press, 2014:1636-1642.

[2] Cai Y, Leung H F, Li Q, et al. Typicality-Based Collaborative Filtering Recommendation[J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(3):766-779.

[3] Gong S. A Collaborative Filtering Recommendation Algorithm Based on User Clustering and Item Clustering[J]. Journal of Software, 2010, 5(7):745-752.

[4] Hosseini S M S, Maleki A, Gholamian M R. Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty[J]. Expert Systems with Applications, 2010, 37(7):5259-5264.

[5] 赵妍, 赵学民. 基于CURE的用户聚类算法研究[J]. 计算机工程与应用, 2012, 48(11):97-101.

[6] 张文东, 易轶虎. 基于兴趣相似性的Web用户聚类[J]. 山东大学学报(理学版), 2006, 41(3):45-47.

[7] Kaur J, Madan N. Association Rule Mining: A Survey[J]. International Journal of Hybrid Information Technology, 2015, 8.

[8] Jannach D, Zanker M, Felfernig A, et al. Recommender Systems: An Introduction[M]. Cambridge University Press, 2010.

[9] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]// International Conference on World Wide Web. ACM, 2001:285-295.

[10] Tan P N, Steinback M, Kumar V. Introduction to Data Mining[J]. Data Analysis in the Cloud, 2006, 26(25):1-25.

[11] https://stackoverflow.com/questions/35709562/how-to-calculate-clustering-entropy-a-working-example-or-software-code