一种个性化推荐方法

朱 宝[1] 徐玲玉[2]

(华南理工大学自动化科学与工程学院 广州 510640)[1] (合肥工业大学管理学院 合肥 231500)[2]

摘 要 提出了一种新的个性化推荐方法。该方法来源于对个性化推荐技术本质的研究。产出的方法包括一种用正态分布卷积性质所得到的离线相似度计算方法;一种通过计算物品与物品之间无差别的相似性操作次数得到离线相似度的方法;一种用类似于贝叶斯的方法来综合不同的相似度结果的方法。另外还提到一些用于工程实施的方法和技巧。所提方法已经在数据挖掘领域得到了成功的应用。

关键词 个性化推荐,相似性,数据挖掘

中图法分类号 TP391 **文献标识码** A

Personalized Recommendation Technology

ZHU Bao[1] XU Ling-yu[2]

(School of Automation Science and Engineering, South China University of Technology, Guangzhou 510640, China)[1]

(School of Managements, Hefei University of Technology, Hefei 231500, China)[2]

Abstract This paper proposed a new personalized recommendation method. The method is derived from the research on the essence of personalized recommendation technology. We arrived at a solution which includes several methods: one is calculating the offline similarity by the use of normal distribution's convolution theories, one is calculating the offline similarity by calculating the number of similar operations between each pair of items, the last one is integrating the similarity results of different methods by the use of something similar to the Bayesian formula. We also mentioned some other methods and techniques used to engineering implementation. The method proposed in this paper has been successfully applied in the field of data mining.

Keywords Personalized recommendation, Similarity, Data mining

1 引言

当前大数据领域中,个性化推荐技术应用十分普遍。已有的个性化推荐技术,除了传统的协同过滤方法[1]、矩阵分解方法[2],其他的个性化推荐方法也层出不穷,时间较近的且比较典型的有基于概率的随机游走模型(ProbS)[3]和热传导模型(HeatS)[4]。但是在这些方法的实施过程中,遇到了各种各样的问题。本文通过深入研究个性化推荐技术、尤其是无差别操作下的个性化推荐技术的根本原理,提出了新的个性化推荐方法。该方法基于详细的理论推导,易于工程实施,能够最大限度地解决当前推荐技术的问题。无论是面对数据量少的情况,还是数据量大、运算能力不足的情况,本文提到的方法都是可以应对的。

2 技术方法与原理

2.1 无差别相似性操作下的个性化推荐原理

假设存在如表1所列的用户行为数据。

表1 用户和物品之间的关联信息

	item1	item2	item3
user1	1	1	0
user2	1	0	1
user3	2	0	0

如果以上表格的操作行为次数,是指无差别含义下的相似性操作次数,假设用户和物品的属性是一维的值可以描述的,分别假设为 x,y,那么有无别的相似性操作 c 满足式(1)。

$$c_{i,j}=k\cdot\frac{1}{\sqrt{2\pi}\delta_i}e^{(-\frac{(x_i-y_j)^2}{2\delta_i^2})} \tag{1}$$

其中,下标 i 和 j 描述数据对应的坐标位置;k 是比例系数;方差 δ 是由用户自身、环境等因素引起的。假设用户有 m 个,物品有 n 个,那么就有 $(m*n)$ 个方程。又假设对应每个用户的方差 δ 相同,比例系数 k 也相同。那么包括未知的用户属性 x,物品属性 y,以及方差 δ 和比例系数 k,一共有 $(m*n+2)$ 个未知数。

方差可以先试一个常数。再增加一个统计量,如所有的用户属性和商品属性的均值为0。那么方差个数和未知数个数都会是 $(m*n+1)$ 个,这样就可以解出来用户属性和商品属性了。

具体的方差需要通过交叉验证的方式预测,使得预测损失指标最低,来尝试获取。但显然采用这个思路来解决推荐问题,运算量巨大。且该方程组化简得到的含绝对值线性方程组,也是很难求解的。

在推荐的时候我们需要的是对无差别的相似性操作次数的预测。而这可以通过以下方法来计算。

朱 宝(1987—),硕士,主要研究方向为大数据和数字信号处理;徐玲玉(1986—),硕士生,主要研究方向为企业管理。

2.2 利用正态分布卷积性质所得到的相似性计算方法

定义 1 满足式(2)的相似度为变量 x 和变量 y 的 δ 相似度。

$$\mathrm{sim}(x,y,\delta)=\frac{1}{\sqrt{2\pi}\delta}e^{\left(-\frac{(x-y)^2}{2\delta^2}\right)} \tag{2}$$

其中，$\mathrm{sim}(x,y,\delta)$是表示相似度的值，它是可列可加的。这个值是某个方差下的似然。

假设又知道式(3)。

$$\mathrm{sim}(y,u,\delta)=\frac{1}{\sqrt{2\pi}\delta}e^{\left(-\frac{(y-u)^2}{2\delta^2}\right)} \tag{3}$$

那么已知式(2)和式(3)，可以得到式(4)。即正态分布的卷积依然是正态分布。这个是正态分布的性质，也可以自行证明，中间涉及一个利用欧拉公式将其转换为极坐标的证明过程[5]。

$$\begin{aligned}&\mathrm{sim}(x,u,\sqrt{2}\delta)\\&=\int_{-\infty}^{+\infty}\mathrm{sim}(x,y,\delta)\cdot\mathrm{sim}(y,u,\delta)\mathrm{d}y\\&=\frac{1}{\sqrt{2\pi}\sqrt{2}\delta}e^{\left(-\frac{(x-u)^2}{2(\sqrt{2}\delta)^2}\right)}\end{aligned} \tag{4}$$

再由式(4)可以计算得到式(5)。

$$\mathrm{sim}(x,u,\delta)=\frac{\mathrm{sim}^2(x,u,\sqrt{2}\delta)}{\int_{-\infty}^{+\infty}\mathrm{sim}^2((x,u,\sqrt{2}\delta)\mathrm{d}u} \tag{5}$$

式(5)不需计算积分也可以看出。因为平方后化为正态分布时，对应的方差部分减少了。且结果归一化仍然是概率密度，那么它的形式必然满足式(2)的形式。

这样我们对照个性化推荐下的数据。如果物品的属性值是来自总体服从均匀分布的样本，那么用户操作的物品，其属性值满足以用户倾向操作的物品属性值为期望、以某种方差为方差的正态分布。若考虑物品只有一个属性，则物品的属性值 y 与用户倾向操作的物品属性值 x 之间的相似度满足式(2)。一般将用户倾向操作的物品属性值简称为用户属性。再考虑用户的属性值是来自总体服从均匀分布的样本，那么操作物品的用户，其属性值满足以物品的属性值为期望、以某种方差为方差的正态分布。即物品的属性值 y 与用户的属性值 u 之间的相似度满足式(3)，然后利用式(4)便可得到用户的属性 x 和用户的属性 u 之间的相似度关系。再利用式(5)可以得到很多满足定义 1 的相似度。因为物品都可以用独立的属性进行描述，且属性独立，所以同样可以利用正态分布的卷积性质，得到似然，即满足定义 1 下的相似度了。

一般用户和物品的属性值都不是来自总体服从均匀分布的样本。我们再利用下面的方法来逼近满足定义 1 中的相似度。

2.3 利用相似性信息的含义得到的相似性计算方法

首先区别定义相似性和相似度。

定义 2 相似性是指物体和物体之间具有一些相同的属性。

定义 3 相似度是指相似性的度量，是与物体之间所具有的无差别含义的属性的个数。

该相似度定义和信息量的含义相似[6]，但便于计算理解。该相似度是可列可加的。

来自某个未知总体下的样本集合 Ω，由样本 ω 的独立同分布，知道该集合中的样本和样本之间的相似度相同(样本来自于同一未知总体，是样本和样本之间唯一已知的相同的属性)。

记已知的一个无差别含义的属性，为一个相似度单位。样本 ω_i 和样本 ω_j 之间的相似度为：

$$\mathrm{sim}\_\omega(\omega_i,\omega_j)=1(\mu_\omega) \tag{6}$$

其中，变量 i,j 是下标，μ_ω 为相似度单位，与具体相似性属性有关。由于样本是具有属性的，因此具有属性 θ_i 的样本集合 Ω_i 和具有属性 θ_j 的样本集合 Ω_j 的相似度为：

$$\mathrm{sim}\_\omega(\Omega_i,\Omega_j)=num\_\Omega_i\cdot num\_\Omega_j(\mu_\omega) \tag{7}$$

其中，num_Ω_i 和 num_Ω_j 为集合中样本的个数。

当属性 θ_i 和属性 θ_j 是来自未知总体下的“属性样本”的集合 θ 时，有相似度：

$$\mathrm{sim}\_\theta(\theta_i,\theta_j)=1(\mu_\theta) \tag{8}$$

同理 μ_θ 为相似度的单位。现在每发生一个样本，假设其同时具有属性 θ_i 和属性 θ_j，即可理解为，属性 θ_i 和属性 θ_j 是来自了某个同一未知总体下的“属性样本”的集合。那么属性 θ_i 和属性 θ_j 就有了相似度，当这样的样本发生 n 次后，其相似度为：

$$\mathrm{sim}\_\theta(\theta_i,\theta_j)=n(\mu_\theta) \tag{9}$$

在应用中，相似度用于对比，那么相似度单位必须一致。这里相似度单位和具体表示的相似性信息含义有关，所以相似性单位具体表示的相似性信息含量，对于对比结果的正确性有较重要的意义。如利用式(9)中的方法在不同的属性组合上所计算得到的相似度单位信息含量可以更高，因为在利用式(9)进行计算之前，已经知道 $\mathrm{sim}_\theta(\theta_i,\theta_j)$和式(7)有关，即具有属性 θ_i 的样本集合 Ω_i 和具有属性 θ_j 的样本集合 Ω_j，其相似值越大，$\mathrm{sim}_\theta(\theta_i,\theta_j)$越高。可以这么理解：同时具有属性 θ_i 和属性 θ_j 的样本组合可能性越多，那么实际发生的同时具有属性 θ_i 和属性 θ_j 的样本的概率也越大。

所以在更多的相似性信息情况下，我们得到属性 θ_i 和属性 θ_j 的相似度公式：

$$\mathrm{sim}\_\theta(\theta_i,\theta_j)=\frac{n(\mu_\theta)}{num\_\Omega_i\cdot num\_\Omega_j(\mu_\omega)} \tag{10}$$

这样，该相似度用于对比时所产生的错误率更低。通过这样处理的相似性操作次数，就是一种无差别含义下的相似性操作次数。

稍作整理，就有了本文所提出的一般的相似度计算方法。

定义 4 属性集合 θ 中的属性 θ_i 和属性 θ_j 在一个服从未知分布的总体的样本中，观察到属性 θ_i 的发生次数为 $s(i,*)$，观察到属性 θ_j 的发生次数为 $s(*,j)$，以及属性 θ_i 和属性 θ_j 同时发生的次数为 $s(i,j)$，那么有属性 θ_i 和属性 θ_j 的一个相似度估计如式(11)所示。

$$\mathrm{sim}(\theta_i,\theta_j)=\frac{s(i,j)}{s(i,*)\cdot s(*,j)}(\mu) \tag{11}$$

其中，变量 i,j 是自然数，μ 为相似度单位。

定义 5 属性集合 θ 中的属性 θ_i、属性 θ_x 和属性 θ_j 在一个服从未知分布的总体的样本中，利用定义 4 的方法得到估计的相似度 $\mathrm{sim}(\theta_i,\theta_x)$和估计的相似度 $\mathrm{sim}(\theta_x,\theta_j)$，那么属性 θ_i 和属性 θ_j 的一个相似度估计如式(12)所示。

$$\mathrm{sim}'(\theta_i,\theta_j)=\sum_x(\mathrm{sim}(\theta_i,\theta_x)\cdot\mathrm{sim}(\theta_x,\theta_j))(\mu') \tag{12}$$

其中，变量 i,j,x 是下标，μ'为新的相似度单位，即无差别含义下相同属性的个数。

那么现在已知集合 A 和集合 B，以及集合 A 中的元素 a 和集合 B 中的元素 b 的无差别含义下相同属性的个数 $\mathrm{sim}(a,b)$，利用以下公式就可以求取集合 B 内的元素 b_i 和 b_j 之间的相似度。

$$\mathrm{sim}(b_i,b_j)=\sum_a\left(\frac{\mathrm{sim}(a,b_i)*\mathrm{sim}(a,b_j)}{(\sum_b\mathrm{sim}(a,b))^2*\sum_a\mathrm{sim}(a,b_i)*\sum_a\mathrm{sim}(a,b_j)}\right) \quad (13)$$

其中，i 和 j 是下标。因为该相似度是可列可加的，相似度被进行归一化后就可以看成概率。它是与满足定义 1 的在某个方差下的似然成正比的。这里有兴趣的读者可以对比基于概率的随机游走模型(ProbS)[3]和热传导模型(HeatS)[4]，以及其他侧重于给出计算物品和物品之间相似度的方法。这里的相似度是对称的，并且具有更少的假设和更明确的含义。这个在后面提到该相似度时可以很好地进行相似性综合，更可以看出来。

另外某个用户的属性值是随时间变化的。那么对于式(13)，如果集合 A 的元素值随时间变化，那么有式(14)。

$$\mathrm{sim}(b_i,b_j)=\sum_a\left(\frac{\mathrm{sim}(a,b_i)\cdot\mathrm{sim}(a,b_j)\cdot f(t_i,t_j)}{(\sum_b\mathrm{sim}(a,b))^2\cdot\sum_a\mathrm{sim}(a,b_i)\cdot\sum_x\mathrm{sim}(a,b_j)}\right) \quad (14)$$

其中，t_i 和 t_j 分别是元素 a 同元素 b_i 和 b_j 发生关联的时刻点。函数 $f(t_i,t_j)$ 是和这两个时刻点有关的函数。这样的函数根据具体情况确定。如我们只知道元素 a 的值随时间缓慢变化，那么可以将函数定为一个低通滤波函数。

$$f(t_i,t_j)=\beta^{|t_i-t_j|} \quad (15)$$

其中，β 是低通滤波系数。如果我们知道元素 a 是以时长 T 进行周期变化的，那么可以将函数定义为和周期 T 有关的滤波函数[7]。如果这个函数是在时间上有向的，那么可以用该方法寻找数据中的频繁项等。

如果集合 A 和集合 B 是同一个集合，那么利用式(13)中的计算方法计算的结果在归一化之后可以看作是方差增大的情况下的似然。在实际的情况下，方差的增大就会产生更多的相似度关联。这对于需要这个作用的个性化推荐问题，会带来更好的推荐效果。即可以关联到的推荐结果增加了。

2.4 利用类似于贝叶斯公式的方法来综合不同相似性结果的方法

现有算法的一个很大的问题是，计算的相似性结果其含义不明确。那么在利用多个数据源进行相似度计算的时候，各个数据源计算得到的相似度如何综合，就很成问题。以上的相似度结果是可列可加的，其归一化结果，可以看成概率密度。那么就可以采用类似于贝叶斯公式的一种方法，来综合不同数据源下的相似度结果。这里先说和贝叶斯公式一致的方法。假设通过某个数据集计算得到的结果为 $\mathrm{sim}'(b_i,b_j)$，通过另外一个数据集计算得到的结果为 $\mathrm{sim}''(b_i,b_j)$。如果得到 $\mathrm{sim}'(b_i,b_j)$ 考虑的属性比得到 $\mathrm{sim}''(b_i,b_j)$ 考虑的属性少，那么可以将 $\mathrm{sim}'(b_i,b_j)$ 命名为先验概率，将 $\mathrm{sim}''(b_i,b_j)$ 命名为条件概率。然后我们有可以类比贝叶斯公式的式(16)。

$$\mathrm{sim}'''(b_i,b_j)=\frac{\mathrm{sim}'(b_i,b_j)\cdot\mathrm{sim}''(b_j,b_i)}{\sum_j\mathrm{sim}'(b_i,b_j)\cdot\mathrm{sim}''(b_j,b_i)} \quad (16)$$

对考虑属性较少所得到的相似性结果，可以适当地补充白色噪声，即认为，未知的属性信息对相似度结果贡献了白色的噪声。于是我们有式(17)。

$$\mathrm{sim}'''(b_i,b_j)=\frac{(k\cdot\mathrm{sim}'(b_i,b_j)+(1-k)/n))\cdot\mathrm{sim}''(b_j,b_i)}{\sum_j(k\cdot\mathrm{sim}'(b_i,b_j)+(1-k)/n))\cdot\mathrm{sim}''(b_j,b_i)} \quad (17)$$

其中，n 是集合 B 中元素 b 的个数，k 是控制已知属性信息贡献比例的系数。若两组相似度考虑的属性互有交叉，那么可以都增加一部分白色噪声。再考虑到相似度是对称的，且不考虑归一化的问题，那么相似度综合方法实际上可以写成以下形式的。

$$\mathrm{sim}'''(b_i,b_j)=\frac{(\mathrm{sim}'(b_i,b_j)+N_1)\cdot(\mathrm{sim}''(b_i,b_j)+N_2)}{\sum_j(\mathrm{sim}'(b_i,b_j)+N_1)\cdot(\mathrm{sim}''(b_i,b_j)+N_2)} \quad (18)$$

其中，N 是需要补充的噪声。这里由于存在两个噪声参数，在利用交叉验证来调试的时候，稍稍困难一些。另外已知贝叶斯公式中，相似度计算结果和两组相似度的方差有关、方差较小的在计算过程中，对结果贡献更大。这可以通过式(5)和交叉验证的方法进行调整。

综上，利用这种方法，就可以综合不同数据所得到的相似性信息了。

2.5 实际工程实施中的一些技巧补充

具体利用以上方法进行个性化推荐时，为了推荐效果更好，或者为了运算量适当，补充以下的部分。

在用户没有对物品进行操作的时候，为了给用户产生推荐，同时利于平滑过渡，初始情况下，可以认为用户操作了一定次数的物品。操作的次数，以物品被操作的次数为比例，落在不同的物品之上。这样初始情况下就可以为用户推荐物品。但这样同时会大大增加数据量和运算量。可以以物品被操作的次数为概率比例，落在不同的物品之上，或者只取排名靠前的一定数量物品，以物品被操作的次数为比例，落在不同的物品之上。

在利用式(13)和式(16)－式(18)进行相似度关联的增强来增加相似度关联的个数、提升推荐效果的时候，可以依据机器的运算能力，适当地增强相似度关联。这样可以做到推荐效果和算法复杂度之间的最佳折衷。

3 技术实施实例

3.1 数据处理过程举例

假设存在表 1 中的用户和物品之间的相似性操作次数，利用式(11)可以处理得到无差别操作次数，如表 2 所列。

表 2 物品和属性之间的关联信息

	item1	item2	item3
user1	0.1250	0.5000	0
user2	0.1250	0	0.5000
user3	0.2500	0	0

再利用式(12)可以得到物品和物品之间的相似度，如表 3 所列。

表 3 物品和物品之间的相似度

	item1	item2	item3
item1	0.0938	0.0625	0.0625
item2	0.0625	0.2500	0
item3	0.0625	0	0.2500

这和直接利用表 1 中数据，在式(13)下计算的结果是相

同的。假设存在表 4 中物品和属性之间的关联信息。

表 4　物品和属性之间的关联信息

	tag1	tag2	tag3
item1	1	1	0
item2	1	0	1
item3	1	0	0

这里直接用式(13)就可以得到物品和物品的相似度，如表 5 所列。

表 5　物品和物品之间的相似度

	item1	item2	item3
item1	0.2778	0.0278	0.0556
item2	0.0278	0.2778	0.0556
item3	0.0556	0.0556	0.1111

在计算资源允许的情况下，有时候我们还对以上的相似度进行增强。如我们对表 3 中的相似性结果，利用式(13)进行增强，那么有表 6 的相似性结果。

表 6　物品和物品之间增强的相似度

	item1	item2	item3
item1	0.0174	0.0217	0.0217
item2	0.0217	0.0664	0.0039
item3	0.0217	0.0039	0.0664

由表 6 可以看到，原来稀疏的矩阵被补充得更满了。这在推荐领域是很有用的，可以在一定程度上克服矩阵稀疏的问题。

现在我们对表 6 中的相似度和表 5 中的相似度结果进行综合。这是通过两组不同数据计算得到的相似度。首先我们构造一份噪声数据，如表 7 所列。

表 7　物品和物品之间的噪声相似度

	item1	item2	item3
item1	0.1111	0.1111	0.1111
item2	0.1111	0.1111	0.1111
item3	0.1111	0.1111	0.1111

利用式(18)，这里可以简单地先假设 N_1 的值为 1，N_2 的值都是 10，以表示表 2 中考虑的物品属性相对表 1 中要少。并假设认为表 5 计算得到的相似度和表 6 中相似度的方差相同，不同的情况可以利用式(5)进行调整，调整以最小化某个指标进行交叉验证。那么得到综合的相似度结果如表 8 所列。

表 8　物品和物品之间的相似度

	item1	item2	item3
item1	0.1785	0.1513	0.1550
item2	0.1513	0.2466	0.1342
item3	0.1550	0.1342	0.2170

注意这里是矩阵各个对应元素相乘。利用表 2 中的结果，和表 8 中的结果，并利用式(12)，我们得到表 9 所列的推荐结果。

表 9　物品和物品之间的相似度

	item1	item2	item3
item1	0.4253	0.2838	0.2909
item2	0.2767	0.6608	0.0625
item3	0.3012	0.0664	0.6324

将相似度结果和表 1 中的信息进行综合，就得到如下的最终的推荐结果，如表 10 所列。

表 10　用户和物品之间的关联信息

	item1	item2	item3
user1	0.0980	0.1422	0.0865
user2	0.0998	0.0860	0.1279
user3	0.0446	0.0378	0.0388

这样我们就有了最终的推荐结果。当还有测试数据集的时候，就可以通过测试数据集和交叉验证的方法来使得某个损失函数指标最小，来调节以上的各个参数，以及方差。

如我们可以将表 10 和测试数据集，都针对用户进行归一化。然后针对每个用户的两份概率密度函数，求累计均方根误差。再在所有用户上做平均。将这个结果做个指标，利用交叉验证的方法，使得以上的参数结果最小化。那么就可以定下以上的参数和方差了。

3.2　工程实施情况介绍

该技术现在已经在企业内部得到成功的应用。具体的实施过程中使用 hadoop 离线计算物品的相似结果，再计算推荐结果。推荐效果在某个书城平台上实施，点击 pv 和展示 pv 的比例，提升效果是其他团队利用传统的协同过滤、矩阵分解等方法达到的提升效果的 4 倍以上。其实多少倍，看具体的数据。因为该方法基于详细的推导，计算结果是可以非常接近已有数据集和运算能力上的效果上限的。另外该算法也可以在损失一定信息的情况下，实时计算。

结束语　一般企业中，个性化推荐系统可以获取的信息是行为数据、离散的标签数据，但有的时候还有连续的标签数据。其本质是多类别的模式识别算法。区别于多类别的模式识别算法，个性化推荐算法主要考虑前两种数据。而传统意义上的模式识别主要考虑最后一种数据。

本文所提到的个性化推荐方法，充分地利用了数据中的信息量，并能够使运算复杂度和推荐效果达到很好的折衷。可以综合不同的数据信息，最终达到已有条件下的逼近上限的推荐效果。并且该方法中，如相似度计算的一般方法在模式识别、频繁项挖掘上可以应用。所以该方法还有很强的推广能力。由于本文中的思想方法在个性化推荐领域应经得到了完整的实现，其他领域的应用尚未实现，故暂时仅以此作为本文的主要内容。

参 考 文 献

[1] Goldberg D, Nichols D, Oki B M, et al. Using collaborative filtering to weave an information apestry[J]. Communications of the ACM, 1992, 35(12): 61-70

[2] Lee D D, Seung H S. Learning the Parts of Objects with Non-Negative Matrix Factorization[J]. Nature, 1999, 401: 788-791

[3] Zhou T, Ren J, et al. Bipartite network projection and personal recommendation[J]. Physical Review E, 2007, 76(4): 046115

[4] Zhou T, Kuscsik Z, Liu J G, et al. Solving the apparent diversity-accuracy dilemma of recommender systems[J]. Proceedings of the National Academy of Sciences, 2010, 107(10): 4511-4515

[5] 盛骤，谢式千，潘承毅然. 概率论与数理统计(第四版)[M]. 北京：高等教育出版社，2008

[6] 范明，柴玉梅，昝红英. 统计学习基础[M]. 北京：电子工业出版社，2004

[7] 米特拉. 数字信号处理：基于计算机的方法(第四版)[M]. 北京：电子工业出版社，2012