一种优化的基于用户聚类的过滤推荐策略①

Filtering Recommendation of Optimized Stretegy Based On User Clustering

张海荣　朱信忠　赵建民　徐慧英　(浙江师范大学 数理与信息工程学院 浙江 金华 321004)

摘　要： 本文针对人们对推荐系统的精确性和实时性要求的不断提高的情况,在已有用户聚类方法上进行了优化,提出将聚类标准由实际评分值转换为用户对项目真实兴趣度,提高聚类的精确性,根据用户兴趣相似的特征改进计算用户相似性的方法。

关键词： 协同过滤 最近邻居 聚类 兴趣度 相似

近年来,随着网络的大量普及,各种各样的信息充斥着网络的每一个角落。不同的人不同时期对网络信息有不同的需求,这种情况下信息过滤推荐系统越来越受到人们的重视。推荐系统中最近邻协同过滤推荐技术是当前最成功的推荐技术[1],它的主要思想是根据用户对项目评分的相似性查询出最近邻居集合,根据集合中的评分数据向目标用户产生推荐。

为了产生更加精确的推荐,保证推荐系统的实时性要求,研究人员提出了基于聚类技术的推荐算法技术,根据用户对项目的评分,将评分数值相似的用户分配到相同的簇中,当目标用户出现时首先判断目标用户所在的簇,根据簇中其他用户对商品的评分预测目标对该商品的评分。由于聚类过程可离线进行,所以在线的推荐算法产生推荐的速度比较快。

然而这些聚类技术中的聚类标准均为用户对项目的评分值,这种聚类标准有很大的局限性。因为每个用户的评分标准不同,例如用户 U_1 习惯把最喜爱的商品评为 5 分,而用户 U_2 习惯把最喜爱的商品评为 10 分,如果 U_1、U_2 对同一商品均评了分值 2 分,显然从分值的意义看两个用户对该商品的喜爱程度是不同的,U_1、对商品的喜爱度高于 U_2,但是从分值上看却是相同的,现有的聚类方法必然将他们放在同一个聚类中,这使聚类非常不精确,产生了误差。本文针对这样的局限性,提出一种将用户评分转换为对项目感兴趣程度的思想,用于在同一水平挖掘用户对商品真实的兴趣度,用这样值替换实际评分值作为聚类标准进行聚类,就大大增加了精确性。

1　优化的基于用户聚类的过滤推荐策略介绍

用户评分数据可以用一个 m * n 阶 A(m,n)表示,m 行代表 m 个用户,n 列代表 n 个项目。$R_{i,j}$ 代表第 i 行第 j 列的元素,即为用户 i 对项目 j 的评分。用户－项目矩阵如表 1 所示[2]:

表 1　用户———项目矩阵

	$Item_1$	.....	$Item_j$	.....	$Item_n$
$User_1$	$R_{1,1}$	.....	$R_{1,k}$	.....	$R_{1,j}$
.....	.....	.....	.....	.....	.....
$User_i$	$R_{i,1}$	.....	$R_{i,j}$	.....	$R_{i,n}$
.....	.....	.....	.....	.....	.....
$User_m$	$R_{m,1}$	.....	$R_{m,j}$	.....	$R_{m,n}$

用户对某个项目有评分则矩阵中就有相应的评分值 $R_{i,j}$,否则便默认为 0. 一个注册用户是否对该网站的项目进行评分完全取决于用户自身的意愿,如果建

① 基金项目:国家自然科学基金项目:(60773197);浙江省自然科学基金:(Y107750)

立所有用户和所有项目对应的用户－项目矩阵显然是不合适的[3]。没有参与评分的用户和没有被任何用户评分过的项目是没有任何意义的，并且还会误导系统认为用户对该项目不感兴趣，评分为 0。对于用户－项目矩阵来说，这些无意义的即是所有横向值为 0 的用户和所有纵向值为 0 的项目，将这些用户和项目去掉后的矩阵相对比较有价值。

即使如此优化后矩阵还是有些稀疏，因为每一位用户不可能对所有的项目均做出了评分，矩阵中仍然会有值为的 0 的情况。为此本文用文献[4]中提到的根据最近邻居的预测方法预测用户对未评分项目的预测评分，将矩阵填充。计算用户 i 和 j 评分项目集合的并集，用户 i 和 j 在项目并集中未评分的项目通过用户对相似项目的评分加以预测。

$$R_{u,i}=\begin{cases}R_{u,i}, if\ use.u.rated.item.i\\ P_{u,i}, otherwise\end{cases} \quad (1)$$

其中，$R_{u,i}$ 是用户 u 对项目 i 的真实评分，$P_{u,i}$ 是用户 U 对未评分项目 i 的预测评分。

1.1 评分数值转换

单单根据评分数值大小的相近进行聚类，自然会产生不可避免的误差，因为每一位用户的评分标准不一样，不同标准下不同用户对某一项目评分值的异同并不能代表用户对该项目喜好程度的异同。为了使聚类更加精确，本文将用户对项目的实际评分值加以转换，挖掘出用户对项目真实的兴趣度来作为聚类标准进行聚类，提高了聚类的精确性。

m 和 n 分别是一个用户和一个项目，$A_m(n)$ 代表用户 m 对项目 n 的实际评分值，$D_m(n)$ 代表用户 m 对项目 n 的兴趣度。由于用户评分习惯的不同，不同用户对同一项目不同的评分值可能会暗藏着相同的兴趣度。

许多实际问题的大量随机变量都近似服从以随机变量均值和方差为参数的正态分布，本文中用户对项目的评分也满足正态分布[5]，即：

$$A_m(n)\sim N(\mu(A_m(n)),\sigma(A_m(n),m)) \quad (2)$$

其中，$\mu(A_m(n_i))$ 是用户对项目评分的均值，$\sigma(A_m(n),m)$ 是用户对项目评分的方差。

转换步骤：

1. 计算某一用户 m 的平均评分值

$$\mu(A_m(n))=\frac{\sum_{i\in I}A_m(n_i)}{I_n} \quad (3)$$

其中，$A_m(n_i)$ 是用户 m 对不同项目的评分值，I 为项目集合，I_n 为项目集合的元素个数。

2. 计算方差

$$\sigma(A_m(n),m)=\sqrt{\frac{\sum_{i\in I}A_m(n_i)-\mu(A_m(n))}{I_n}} \quad (4)$$

3. 根据用户 m 对项目的评分值 $A_m(n)$ 的概率密度函数求解 m 在不同项目上评分的概率分布，此概率分布即可反映用户对不同项目的兴趣度。

$$\begin{aligned}D_m(n)&=\int_{-\infty}^{A_m(n)}\frac{1}{\sqrt{2\Pi}\sigma(A_m(n),m)}e^{\frac{(t-\mu(A_m(n))^2}{2(\sigma(A_m(n),m))^2}}dt\\&=\int_{-\infty}^{\frac{A_m(n)-\mu(A_m(n)}{\sigma(A_m(n),m)}}\frac{1}{\sqrt{2\Pi}\sigma(A_m(n),m)}e^{\frac{x^2}{2}}\sigma(A_m(n),m)dx\\&=\int_{-\infty}^{\frac{A_m(n)-\mu(A_m(n)}{\sigma(A_m(n),m)}}\frac{1}{\sqrt{2\Pi}}e^{\frac{x^2}{2}}dx\end{aligned} \quad (5)$$

1.2 优化的基于用户聚类的推荐系统

这种过滤推荐的思想是根据用户的兴趣度数据进行聚类，将相近兴趣度的 k 个用户归为一类，由此聚成不同的类。目标用户到达时先判断所属的聚类，然后再相应的聚类中搜索最近邻居，根据最近邻居的兴趣度产生推荐。

1.2.1 聚类算法

将数据库中的对象分类是数据挖掘的基本操作。为了找到效率高、通用性强的聚类方法，人们从不同角度提出了多种聚类方法，其中最典型的有：K－means 算法、K－medoids 算法、CLARANS 算法、CURE 算法等。本文采用分层聚类算法 CURE 算法将用户对项目的兴趣度进行聚类。

设 k 是 CURE 算法的输入参数，代表该方法分割的聚类个数。数据集由 n 个数据点组成，在初始化时，每一个点都是一个初始类，然后将最相近的数据点合并为一个聚类，聚类个数自加，直到聚类中的个数达到参数 k。

CURE 算法将传统对类的表示方法进行了改进。回避了用所有点或中心和半径来表示一个类，而是从每一个类中抽取固定数量、分布较好的点作为描述此类的代表点。

1.2.2 查询最近邻居

当目标用户到达时首先判断它所属的聚类，在目标聚类中查寻与目标用户相似系数最大的若干个用户。

传统计算用户相似性的方法主要有两种:余弦相似性和相关相似性。

余弦相似性:设用户 m_1,m_2 在项目空间上的评分分别表示为向量 m_1,m_2,则两者之间的相似性为:

$$sim(m_1,m_2)=\cos(m_1,m_2)=\frac{m_1*m_2}{|m_1|*|m_2|} \tag{6}$$

其中分子为用户评分向量的内积,分母为用户向量模的乘积。

相关相似性:设用户 m_1,m_2 共同评分过的项目集合为 $I_{m1,m2}$,则两者之间的相似性为:

$$sim(m_1,m_2)=\frac{\sum_{n\in I_{m_1,m2}}(A_{m_1}(n)-\overline{A_{m_1}})(A_{m_2}(n)-\overline{A_{m2}})}{\sqrt{\sum_{n\in I_{m_1,m2}}(A_{m_1}(n)-\overline{A_{m_1}})^2}*\sqrt{\sum_{n\in I_{m_1,m2}}(A_{m_2}(n)-\overline{A_{m2}})^2}} \tag{7}$$

其中,$A_{m1}(n)$ 和 $A_{m2}(n)$ 分别是用户 m_1 和用户 m_2 对某一项目 n 的评分值。

本文对用户的评分值转换为对项目的兴趣度值,这样用户 m_1,m_2 相似性可以从以下几个方面进行描述:

① 在原始用户 - 项矩阵中,用户 m_1,m_2 同时评分过的项目越多,则两者的相似性就越高。

② 用户 m_1,m_2 对同一项目的兴趣度越相似,则两者的相似性就越高。

基于这两个方面,我们提出如下公式来计算用户之间的相似性:

$$sim(m_1,m_2)=\frac{Item(m_1,m_2)}{Item(m_2)+\sum|p_{m_1,n}-p_{m_2,n}|} \tag{8}$$

其中,$Item(m_1,m_2)$ 表示用户 m_1,m_2 共同评分过的项目数,$Item(m_2)$ 表示用户 m_2 评分过的项目数,$p_{m1,n}$ 表示用户 m_1 对项目 n 的兴趣度值。

1.2.3 对目标用户产生推荐

最近邻居产生后,根据最近邻居的兴趣度就可以计算目标用户对项目的兴趣度,产生 Top - N 推荐。设用户 m 和已有的项目集 I,则对任意未评分的项的兴趣度如下[6]:

$$P=\overline{A(m)}+\frac{\sum_{i=1}^{n}corr_i*rating_i-\bar{i}}{\sum_{i=1}^{n}corr} \tag{9}$$

1.2.4 度量标准

评价推荐系统质量的度量标准主要有两种:统计精度度量方法和决策支持度量方法。本文采用统计精度度量方法中的平均偏差法:MAE(mean absolute error)。MAR 通过计算预测用户评分和用户实际评分之间的偏差来度量推荐系统的质量。MAE 越小,推荐质量越高。设预测的用户评分集合为:$E_F=\{p1,p2,p3,...\}$,实际用户评分集合为:$E_P=\{q1,q2,q3,...\}$,则平均绝对偏差为:

$$MAE=\frac{\sum_{i=1}^{N}|p_i-q_i|}{N} \tag{10}$$

1.2.5 试验结果分析

实验所用数据采用浙江师范大学校园网电影 FTP 里的数据,截至目前该网站已注册用户 3058593 人,被评价的电影有 2000 余部。该实验从中选取 4000 余条评分数据,包含 230 个用户和 786 部电影。

根据网站提供的数据集,分别采用传统的原始算法和本文提出的优化算法进行比较,试验结果如下:

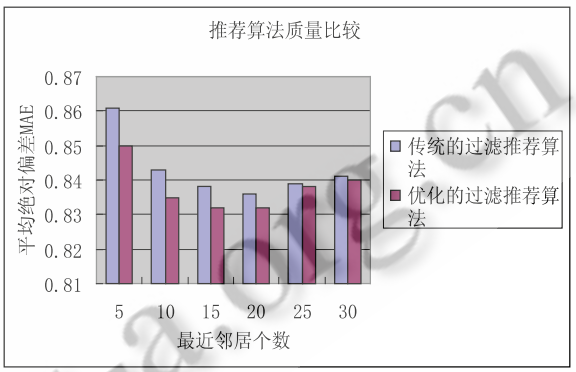

图 1　两种算法的 MAE

从图图 1 两种算法的 MAE 中,可以看到本文提出的优化后的基于用户聚类的过滤推荐算法明显优于传统的过滤推荐算法。将用户聚类的标准转换为用户对项目潜在的兴趣度,加强了聚类的精确性。本文根据用户间相似时表现出的两个特征提出的用户间的相似性公式也增强了查询最近邻居的精确性。

2　结论

随着网络的的发展和普及,人们对网络会越来越依赖,推荐系统规模也会越来越大,人们对推荐系统的高效性和实时性要求越来越高。本文首先挖掘用户对项目潜在的真实的兴趣度,以此为标准对用户进行聚

(下转第 71 页)

(上接第 97 页)

类,然后基于用户相似特征抽象出用户相似性计算公式来寻找目标用户的最近邻居。这种方法大大提高了推荐系统推荐项目的精确性。在计算用户相似性时既要考虑原始用户 - 项矩阵中的原始数据又要考虑转换后的兴趣度,计算起来有些麻烦,这些问题有待于在以后的研究里改进。

参考文献

1 Breese, J Hecherman, D Kadie. C Emirical analysis of predictive algorithms for collaborative filtering. in: Proc of the 14th Conf on Uncertainty in Artificial Intelligence (UAI98). San Francisco, CA: Morgan Kaufmann. 1998. 43 - 52.

2 曾艳,麦永浩. 基于内容预测和项目评分的协同过滤推荐. 计算机应用,2004,24(1):111 - 114.

3 周军锋,汤显,郭景峰. 一种优化的协同过滤推荐算法. 计算机研究与发展,2004,41(10):1842 - 1847.

4 Content - Boosted Collaborative Filtering for Improved Recommendations. Proceedings of the Eighteenth National Conference on Artificial Intelligence, Canada, 2002. 187 - 192.

5 陈晓红,沈洁,顾天竺等. 基于用户潜在偏好的协同过滤. 计算机工程,2007,33(4):42 - 44.

6 王辉,高利军,王听忠. 个性化服务中基于用户聚类的协同过滤推荐. 计算机应用,2007,27(5):1225 - 1227.

7 Sarwar B, Karypis G, Konstan J. Item - Based collaborative filtering recommendation algorithms. In: Proccedings of the 10th International World Wide Web Conference. 2001. 285 - 295.

8 邓爱林,朱扬勇,施伯乐. 基于项目评分预测的协同过滤推荐算法. 软件学报,2003,14(9):1621 - 1628.