个性化推荐中一种新的相似性计算方法

A New Similarity Method Used in Personalized Recommendation

龚松杰 （浙江工商职业技术学院 信息工程学院 浙江宁波 315012）

摘 要： 个性化推荐是电子商务系统中最重要的技术之一，随着用户数目和商品数目的日益增加，在整个商品空间上用户评分数据极端稀疏，传统的相似性度量方法存在各自的弊端，导致推荐系统的推荐质量急剧下降。针对传统相似性度量方法的不足，提出了一种新的基于模糊相似优先比的相似性度量方法，根据项目之间的相似性预测用户对未评分项目的评分，在此基础上，采用相似优先比计算目标用户的最近邻居。实验表明，该度量方式能够提高个性化推荐系统的推荐质量。

关键词： 个性化推荐 相似性 模糊相似优先比 稀疏性

个性化推荐系统被用来帮助用户在大量的信息中寻找感兴趣的内容，它体现的个性化服务已经成为电子商务领域的一个重要功能。在个性化推荐系统中，协同过滤是当前应用最成功的技术[1]。其思想是基于评分相似的最近邻居的评分数据向目标用户产生推荐。由于最近邻居与目标用户对项目的评分非常相似，因此目标用户对未评分项目的评分可以通过最近邻居对该项目评分的加权平均值逼近。

在协同过滤推荐中，为了对目标用户产生推荐，需要搜索目标用户最近邻居，在此过程中，定义用户之间的相似性成为关键问题之一。目前有余弦相似性，相关相似性和最小平方差等。随着电子商务规模的扩大，用于产生推荐的数据将极端稀疏，用上述度量的产生推荐效果将逐步减弱[2,3]。同时，在使用这些相似性度量的协同过滤系统中，部分用户具有相似的喜好，但由于他们仅对相似的项目进行了评分，而没有评分相同的项目，系统并不将其视为近邻，这就产生了所谓的相似不相同问题。

针对传统相似性度量不足，提出了一种新的基于模糊相似优先比的相似性度量方法，根据项目之间的相似性预测用户对未评分项目的评分，在此基础上，采用模糊相似优先比计算目标用户的最近邻居。该方法在一定程度上缓解了数据稀疏性带来的推荐质量问题，克服相似不相同问题，提高系统的推荐生成质量。实验也表明，本文提出的相似性度量要优于目前使用的其他相似性计算方法。

1 相关技术及其分析

1.1 问题描述

用户评分数据可以用一个 m * n 阶矩阵 A(m * n)来表示，m 行代表 m 个用户，n 列代表 n 个项目，第 i 行第 j 列元素 Ri,j 代表第 i 个用户对项目 j 的评分。如表 1 所示。

表 1 用户 – 项目矩阵

Item / User	Item1	Item2	… …	Itemn
User1	R1,1	R1,2	… …	R1,n
User2	R2,1	R2,2	… …	R2,n
… …	… …	… …	… …	… …
Userm	Rm,1	Rm,2	… …	Rm,n

1.2 传统的三种相似性度量方法

在计算邻居用户对目标用户的影响时，需要查找目标用户的最近邻居，这就需要度量。用户的兴趣可以用评分向量来表示，相当于表 1 中的某一行。有三种传统的度量方法：

余弦相似性：把用户评分看作是 n 维项目空间上的向量。通过计算两个向量之间的夹角余弦来度量两个用户之间的相似性。计算公式如下：

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

$$sim(i,j)=(\sum_{k=1}^{n}R_{i,k}*R_{j,k})/(\sqrt{\sum_{k=1}^{n}R_{i,k}^{2}*\sum_{k=1}^{n}R_{j,k}^{2}})$$

Ri,k,Rj,k:用户 i,j 对项目 k 的评分。

相关相似性:通过 Peason 相关系数来度量两个用户的相似性。计算时,首先找到两个用户共同评分过的项目集 Ii,j,然后计算这两个向量的相关系数。计算公式如下:

$$sim(i,j)=(\sum_{c\in I_{i,j}}(R_{i,c}-A_i)(R_{j,c}-A_j))/\sqrt{\sum_{c\in I_{i,j}}(R_{i,c}-A_i)^2*\sum_{c\in I_{i,j}}(R_{j,c}-A_j)^2}$$

Ii,j:用户 i 和 j 共同评分过的项目集,Ri,c:用户 i 对项目 c 的评分,Ai:用户 i 对资源的平均评分。

修正的余弦相似性:在余弦相似性中没有考虑不同用户的评分尺度问题。修正的余弦相似性通过减去项目的平均评分来弥补这种不足,计算公式如下:

$$sim(i,j)=(\sum_{c\in I_{i,j}}(R_{i,c}-A_c)(R_{j,c}-A_c))/\sqrt{\sum_{c\in I_{i,j}}(R_{i,c}-A_c)^2*\sum_{c\in I_{i,j}}(R_{j,c}-A_c)^2}$$

Ac:项目 c 的平均评分。

1.3 传统的三种相似性度量的不足

在余弦和修正的余弦相似性度量方法中,用户没有评分的项目均将评分假设为 0,这样做的好处是可以有效地提高计算性能,但在用户评分数目极端稀疏的情况下,上述假设的可信度并不高,因为用户对所有未评分项目的评分均假设为 0。事实上,用户对未评分商品的喜好程度不可能完全相同,对这些项目的评分也不可能完全相同,由此可见,这两种度量方法并不能有效地度量用户之间的相似性。

在相关相似性度量方法中,先要计算用户 i 和 j 共同评分过的项目集,但是用户只有在比较多的项目上评分比较相似,他们之间的相似性才比较高。在用户评分数据极端稀疏的情况下,经两个用户共同评分的项目集合很小,这样就算计算出他们之间相似度很高,但实际上不一定就很高,因此,相关相似性度量方法也存在一定的弊端。

因此,传统的度量方法在用户评分数据极端稀疏的情况下并不能有效地度量用户间的相似性,从而使得计算出来的最近邻居不准确,导致整个推荐系统的推荐质量急剧下降。

2 基于相似优先比的相似性计算并产生推荐

由于用户 – 项目矩阵的极端稀疏性,本文先根据项目之间的相似性来进行用户没有评价项目的预测,从而形成密集的用户 – 项目矩阵。

2.1 基于模糊相似优先比的相似性计算

根据密集的用户 – 项目矩阵 R = (Rij) m×n,本文采用模糊相似优先比的方法来计算用户之间的相似性。设用户集合 U = { U1,U2,…,Um },项目集合 I = { I1,I2,…,In }。在密集的用户 – 项目矩阵中选定其中一个测试用户 Uk (1 ≤k ≤m),在用户集合中,找出哪些与 Uk 最相似,哪些次之。具体的方法是以成对的用户(Ui , Uj) ∈U 就 Ip ∈I(1 ≤p ≤n) 项目因素跟测试用户 Uk 比较。当 Ui 与 Uk 的相似程度高于 Uj 跟 Uk 相似程度时,优先选择相似程度大的 Ui,或者说 Ui 比 Uj 优先。

就某一项目因素 Ip ∈I(1≤p≤n),把任意(Ui , Uj) ∈U 跟 Uk 作比较(i , j = 1 ,2 , … , n),得到模糊相似优先关系矩阵

C = (Cij) m×n (Cij ∈\[0 ,1 \] i , j = 1 ,2 , □, n)

其中:Cij 采用绝对值距离,计算公式如下:

$$C_{ij}=(|U_k-U_j|)/(|U_k-U_i|+|U_k-U_j|)$$

$$C_{ji}=(|U_k-U_i|)/(|U_k-U_i|+|U_k-U_j|)$$

显然 Cij + Cji = 1,也符合(Ui , Uj)模糊相似优先比互补的特性。

根据 Cij 来确定 Ui 和 Uj 哪个优先,如果 Cij 在(0.5,1)之间,表示 Ui 较 Uj 优先;如果 Cij 在(0,0.5)之间,表示 Uj 比 Ui 优先;如果 Cij = 1,表示 Ui 绝对比 Uj 优先;如果 Cij = 0.5,表示 Ui 与 Uj 等价;如果 Cij = 0,表示 Uj 绝对比 Ui 优先。

通过对模糊相似优先关系矩阵 C m×n 的 λ 水平评出相似程度,可以得到在项目因素 Ip 的其他用户跟目标用户 Uk 的相似性程度排序。记为 Ip(U1,U2,…,Um) = (λ1,λ2,…,λm),其中 λi 是用户 Ui 的 λ 水平评出相似程度,具体为 Mi = min{ Cij },j = 1 ~ m ; λ = max{ Mi },i = 1 ~ n。同样,通过对其他 Ij ∈I 的模糊相似优先运算,最后得到 Ui 的用户 – 项目 λ 水平评出相似程度矩阵。

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

表 2 用户 – 项目 λ 水平评出相似程度矩阵

User \ Item	I1	I2	… …	In
U1	λ1,1	λ1,2	… …	λ1,n
U2	λ2,1	λ2,2	… …	λ2,n
… …	… …	… …	… …	… …
Um	λm,1	λm,2	… …	λm,n

用户 Uk 跟任一用户 Ui 的相似性为

$$Sim(U_k, U_i) = (\sum_{j=1}^{n} \lambda_{ij}) / n$$

2.2 产生推荐

计算目标用户 U 对未评分项目 I 的评分时，采用下式计算 U 对 I 的评分：

$$P_{ui} = A_u + (\sum_{m=1}^{n} (R_{mi} - A_m) * sim(u,m)) / (\sum_{m=1}^{c} sim(u,m))$$

Au：用户 u 对资源的平均评分，Rmi：用户 m 对项目 i 的评分，Am：用户 m 对资源的平均评分，sim（u，m）：用户 u 和 m 的相似度。

3 实验及结果分析

3.1 数据集和推荐精度度量标准

实验采用的数据集是 MovieLens。它是基于 Web 的研究性的推荐系统，在 1997 年建立。MovieLens 数据集包含 movies. dat、ratings. dat 和 users. dat。movies. dat 中包含了 1682 部电影的详细描述信息，users. dat 中包含 943 位用户的详细信息，ratings. dat 中包含 943 位用户对 1682 部电影的 100000 条评分记录，评分值为从 1 到 5 的整数。

利用平均绝对误差 MAE 来衡量算法的预测精度。MAE（Mean Absolute Error）是测试集中所有用户对资源打分的实际值与预测值的偏差的绝对值的平均[4]。MAE 较早的在 Shardanand & Mases 及 Sarwar 等中用于评价系统预测的性能。MAE 值越小说明推荐算法的预测精度越高。

3.2 推荐精度试验

在基于模糊相似优先的相似性度量的推荐方式中，把与目标用户相似性最高的前 n 个用户作为目标用户的最近邻居。而最近邻居的多少影响着系统的推荐精度。实验中，我们把它与传统的推荐算法进行了比较，实验结果如下。

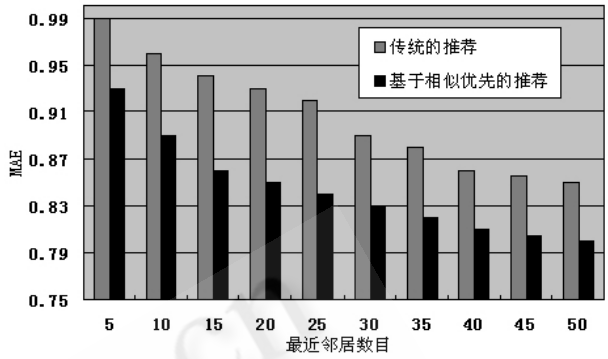

图 1 推荐算法推荐精度比较

从对比实验可以看出，在相同条件下，基于模糊相似优先比的相似性度量的协同过滤算法的 MAE 小于传统的协同过滤算法。由于传统的相似性计算方法仅仅考虑用户对各个项目的实际评分，并没有将项目之间的关系考虑进去，这将影响推荐系统中最近邻居的查找，进而影响推荐结果质量。

4 总结

针对传统的相似性度量方法存在各自的弊端，提出了一种新的基于模糊相似优先比的相似性度量方法，根据项目之间的相似性预测用户对未评分项目的评分，在此基础上，采用相似优先比计算目标用户的最近邻居。实验结果表明，该度量方式能够提高推荐系统的推荐质量。

参考文献

1 Herlocker J, Konstan J, Terveen L, Riedl J. Evaluating collaborative filtering recommender systems. ACM Trans. on Information Systems (TOIS), 2004,22(1):5 – 53.

2 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.

3 Sarwar B, Karypis G. Item – based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International world wide web conference. 2001: 285 – 295.

4 陈健，印鉴. 基于影响集的协作过滤推荐算法. 软件学报，2007，18(7):1685 – 1694.

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn