一种基于分类的协同过滤算法

A collaborative filtering algorithm based on classify

徐义峰　（衢州学院　浙江衢州 324000）
陈春明　（桂林电子工业学院图书馆 广西桂林 541004）
徐云青　（衢州学院　浙江衢州 324000）

摘要：协同过滤技术是当前研究的热点。本文简单地介绍了基于最近近邻法协同过滤算法，针对其不足，提出了一种基于分类的协同过滤算法，并在算法中引入用户权威性来衡量用户评价资源客观性和准确性，使用户推荐更符合“邻居”的需求，进而增强协同过滤推荐资源的准确性。

关键词：协同过滤　分类　算法　用户权威性

1　引言

协同过滤技术试图通过跟踪用户的行为，自动分析大量用户之间的兴趣相似程度，发现和特定用户（目标用户）兴趣最相似的用户群体，然后利用与目标用户相似的用户群体的意见向目标用户提供可能感兴趣的信息。协同过滤技术是根据用户之间的相似性来互相推荐资源的技术，是保证个性化信息服务准确性的基本技术。协同过滤的优点是能够应用于那些不能直接获取特征的资源，还能为用户发现更多新的兴趣[1]。

基于最近近邻的协同过滤算法[2]是根据目标用户以往的评分和最近近邻的评价，预测目标用户对集合中目标用户尚未评价过的信息资源的预测值。该算法生成一个用户的信息资源推荐列表的主要步骤：

（1）计算每个用户的平均打分；

（2）计算用户与用户之间的相似度；

（3）生成预测结果。基于最近近邻法的协同过滤算法的特点是具有很好的预测精度，但在使用初期由于系统资源还未获得足够多的评价，很难发挥作用，即存在数据稀疏性问题；另外该算法还能随数据的变化而变化，比较适合于数据更新频繁的系统，但是随着用户和资源的增多，系统的性能会越来越低，即可扩展性较差。

事实上，对于大部分用户尤其是学术用户来说，他们的兴趣往往集中在某个（或几个）领域，所以用户对信息的评价一般会出现集中在某几个类别中。将资源划分为不同的类型，用户只对一类（或几类）资源感兴趣，所以资源数目和用户数目都将大大降低，这样就可以克服高稀疏性和系统可扩展不好的问题，故本文提出一种基于分类的协同过滤方法。

2　基于分类的协同过滤模型

2.1　资源的原始评价值

在传统的协同过滤算法中，只有被访问过的资源才能被推荐给相似用户，没有访问过的资源很难得到推荐。因此，我们可以通过资源本身的特征给每个资源赋以“原始评价值”，使所有资源都能得以推荐。然后通过用户的反馈来改变这个原始评价值。资源的“原始评价值”可以根据资源的一些特殊标记来表现，例如资源是否是基金项目，资源由著名人物发表，资源的出处等等。

2.2　资源和用户分类

把资源按照某种分类标准（如中图法等）分在不同的资源类别中，如果属于交叉学科知识，分别分在两个不同的类别中。在此模型中，假定资源类别的集合为 $C=\{c_1,c_2,\cdots\cdots,c_k\}$，其中 k 为模型的大小，c_k 表示第 k 个资源类别，则资源表示为一个条件概率的矢量：$Q_i=<p(c_1\mid d_i),p(c_2\mid d_i),\cdots\cdots,p(c_k\mid d_i)>$，其中 d_i 代表资源 i 的加权关键词向量，$p(c_j\mid d_i)$ 表示资源 i 在类别 c_j 中的概率，用 q_{ij} 表示。

资源矩阵表示为：

$$Q=\begin{bmatrix}Q_1\\Q_2\\\vdots\\Q_n\end{bmatrix}=\begin{bmatrix}q_{11}&q_{12}&\cdots&q_{1k}\\q_{21}&q_{22}&\cdots&q_{2k}\\\vdots&\vdots&\vdots&\vdots\\q_{n1}&q_{n2}&\cdots&q_{nk}\end{bmatrix}$$

对用户的分类建立在用户对资源的评分数据和资源特征的基础上。首先获知用户对哪些资源打过分，再了解用户打过分的这些资源分别属于哪几个类别，然后把用户兴趣划分到不同的类别中。一般地，用户关注的资源局限于一定的领域内，即在所有信息资源中，用户对资源的访问和评价集中在几个类别里，很少有可能浏览所有类别的信息资源。如果用户对某一类别中的资源没有评分或用户没有此类别的知识背景，说明该用户对此类资源不感兴趣，所以此类不在推荐行列，这样可降低运算的复杂性。

2.3 用户权威性

传统的最近近邻算法只考虑了近邻的相似性而忽略了近邻的权威性。不同用户对资源的评价可靠程度不同，有经验的用户或权威用户能够给资源更准确、更客观的评价。可见，用户对资源评价的可靠程度（即用户的权威性），反映了用户评价的稳定性。因此权威用户对资源的评价更值得其他用户参考。权威性的确定：

（1）用户评价过的资源数量。如果一个用户在某一方面评价过很多资源，那么他已经具备一定的经验，并在评价时表现出自己的品味，因此用户权威性的第一个特征可以利用用户对资源的评价数目来反映。

$$W_1=\begin{cases}1, & n_u\geq A\\ \dfrac{n_u}{A}, & n_u<A\end{cases}$$

其中，n_u 表示用户 u 所评价过的资源数目。A 为一个常数，称作惩罚数目，一般取 50。当 $n_u<A$ 时，削弱该用户的权威性，若 $n_u>=A$ 时权重为 1。

（2）用户学术背景。如果一个用户评价的资源不多，用户的学术背景也将影响用户权威性。例如，教授和学生同时从事某个领域的研究，一般情况下教授的评价更具有客观性。所以，用户权威性的第二个特征可以利用用户的学术背景来定义：W_2 为 1（教授）；为 K_1（副教授）；为 K_2（讲师）；为 K_3（学生）等。

以上两个方面都是从用户自身角度来衡量的，因此有：$AU_1(u)=\alpha_1 W_1+\alpha_2 W_2$，其中 $\alpha_1+\alpha_2=1$。

（3）从资源的角度来衡量。当一个资源获得足够多的评价时，这个资源评价的平均值可以用来衡量该资源真正的质量或品质，而一个权威用户一般能正确评价事物的本质特征，因此一个用户的评价值接近资源品质的程度，也就是接近资源评价平均值的程度可以用来刻画用户权威性的第三个特征，定义：

$$AU_2(u)=\frac{\sum_{i\in v_u}\left(1-\dfrac{|v_{u,i}-\bar{v}_i|}{Max-Min}\right)}{n_u}$$

其中 v_u 表示用户 u 评价过的资源集合，$v_{u,i}$ 表示用户 u 对资源 i 的评价值，$\bar{v}_i$ 表示资源 i 评价的平均值，Max 和 Min 表示评价的最大值和最小值。

综合用户权威性的三个特征，那么一个用户的权威值可以表示成：$AU(u)=AU_1(u)\times AU_2(u)$。

2.4 相似性计算

（1）资源类别的评价。资源之间有一定的依赖关系，利用这些依赖关系实现相似性计算的特征加权，从而提高协同过滤的精度。属于同一类型的资源存在很强的相关性，因此，可以将资源进行分类。对一个用户来说，可以将评价过的同一类资源的评价值的总和看作是用户对该类资源的评价，这样，用户对资源的评价可以转换为用户对资源类别的评价。

$$v_{a,k}=\sum_{\text{资源}|\text{属于类型}c_k} v_{a,i}$$

其中，$C_k=\{c_1,c_2,\ldots,c_k\}$，K 表示类型的个数，$v_{a,k}$ 表示用户 a 对类 c_k 的评价值，由于类别个数远小于资源个数，因此克服了数据稀疏性带来的问题。

（2）用户相似性计算。基于用户间的相似性计算借用皮尔逊相似性计算方法加上目标资源和资源类型之间的关系，采用改进的相似性计算方法。

如果资源 i 仅属于一类资源则在该资源类别中利用皮尔逊相似性计算方法计算用户间的相似性：

$$W_{a,b}=\frac{\sum_{i=1}^{M}(v_{a,j}-\bar{v}_a)(v_{b,j}-\bar{v}_b)}{\sqrt{\sum(v_{a,j}-\bar{v}_a)^2\sum(v_{b,j}-\bar{v}_b)^2}}$$

其中 $W_{a,b}$ 表示用户 a 与用户 b 的相似性；M 表示资源的个数；$v_{a,i}$ 为用户 a 对资源 j 的评价值；$\bar{v}_a$ 表示用户评价资源的平均值；j 为用户 a 和用户 b 共同打过分的信息资源。

如果资源 i 属于多个类别则利用下列公式计算：

$$W_{a,b}=\frac{\sum_{k=1}^{k}W_k^2(v_{a,k}-\overline{v_a})(v_{b,k}-\overline{v_b})}{\sqrt{\sum_{k=1}^{k}W_k^2(v_{a,k}-\overline{v_a})^2\sum_{k=1}^{k}W_k^2(v_{b,k}-\overline{v_b})^2}}$$

其中，k 是资源类型的个数。W_k 代表 $P(c_k \mid i_l)$，代表资源 i_l 属于类 c_k 的概率，$\overline{v_a}$是用户对所有类别评价的平均值。

2.5 生成推荐结果

由于引入了用户权威性指标来衡量用户评价资源的可靠程度，因此推荐值为：

$$P_{a,k}=\overline{v_a}+\frac{\sum_{b\in U'}(v_{b,l}-\overline{v_b})\times W'_{a,b}\times AU(b)}{\sum_{b\in U'}W'_{a,b}}$$

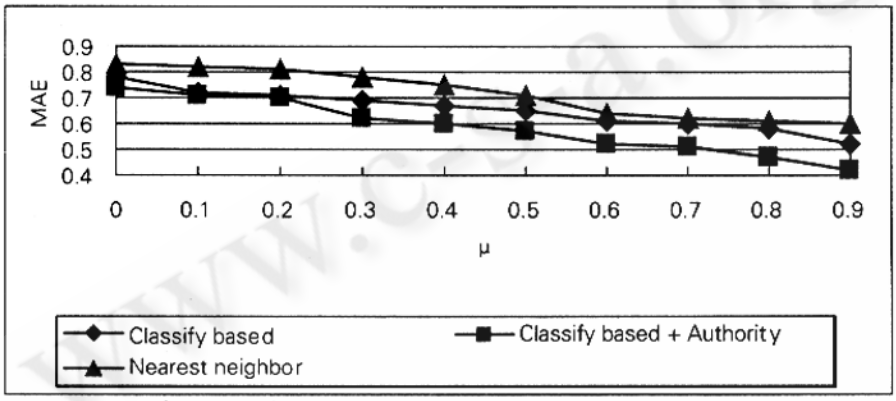

图 1　协同过滤算法准确性比较

其中，$w'_{a,b}=W_{a,b}\times\frac{N_{common}}{50}$，$W'_{a,b}$表示用户 a 和用户 b 相似度的修正值，$N_{common}$表示用户间相交资源的个数，当相交资源的个数大于 50，则影响因子等于 1，采用这个影响因子能提高算法的预测精度，对某项资源，根据公式预测目标用户的打分，选取相似度排名靠前的前 N 个最近邻居参与计算。U' 表示用户 a 的近邻用户集，AU(b) 表示用户的权威性，$\overline{v_a}$表示用户 a 评价过的资源的平均值，$V_{b,l}$表示用户 b 对资源 l 的评价值，$P_{a,k}$表示用户 a 对资源类 c_k 的预测值。

用户 a 对资源 j 的预测值为：$P_{a,l}=W_k * P_{a,k}$。

(1) 如果资源所属多个类别中，综合所有类别的预测值，然后形成资源的推荐值：

$$P_{a,l}=\sum_{k\in U}W_k \cdot P_{a,l}$$

U 为资源所在类别。

(2) 如果用户要求按不同类别分类推荐，则假设用户 a 的兴趣分散在 K 个类别中，推荐的资源项总数为 N，则每一类资源集合中被推荐的是 $P'_{a,l}$排在[N * l_k]前的信息资源。其中：

$l_k=\frac{v_{a,k}}{\sum_{k=1}^{k}v_{a,k}}$，$v_{a,k}$表示用户 a 对 c_k 类资源的打分。

3　实验评估与分析

采用 EachMovie 数据库作为实验数据集，对本文提出的算法与现有的传统协同过滤算法进行比较分析。为了加快实验的进度，我们只从中随机抽取了 5000 个资源评价个数超过 30 的用户，然后将它们分成一个训练集和一个测试集，大约有 38 万个评价值，覆盖所有电影。同时采用平均绝对偏差 MAE (Mean Absolute Error) 作为度量标准验证算法的有效性。平均绝对偏差 MAE 定义为：

$$MAE=\frac{\sum_{x_i\in T}|a_l(x_i)-\hat{a}_l(x_i)|}{card(T)}$$

，其中 $a_l(x_i)$是用户 x_i 对项目 a_l 的实际评分，$\hat{a}_l(x_i)$是用户 x_i 对项目 a_l 的预测评分，T 是测试集，card(T) 表示测试集合的基数。平均绝对偏差 MAE 通过计算用户的预测评分与用户的实际评分之间的偏差，度量预测的准确性。MAE 越小，推荐质量越高。

实验以传统的基于最近近邻法的协同过滤推荐算法为参照，分别计算其平均绝对偏差 MAE；然后与基于分类并引入用户权威性的协同过滤推荐算法作比较，实验结果如图 1 所示。实验结果表明，在各种实验条件下($\mu\in[0,1]$)，与传统的基于最近近邻法的协同过滤推荐算法相比，基于分类的协同过滤推荐算法均具有最小的 MAE，因而其推荐质量较高，也可以看出引入用户权威性能够更客观、更精确的评价资源。

与传统的基于最近近邻法协同过滤推荐算法相比，本文提出的基于分类并引入用户权威性的协同过滤算法在以下两个方面进行了改进。

(1) 运用基于分类近似质量计算用户间的相似性。与传统的基于统计技术的相似性计算方法不同，

本文从分类的角度,将用户的项目评分视作分类知识,在数据稀疏预处理的基础上利用近似分类质量计算两用户的相似性。实验的结果表明,基于分类的协同过滤推荐算法具有最小的平均绝对偏差,推荐质量的最高,有效地克服了数据稀疏对协同过滤推荐质量的影响。

(2) 运用用户的权威性来提高用户对资源评价的可靠程度。传统的最近近邻算法只考虑了近邻的相似性而忽略了近邻的权威性。不同用户对资源的评价可靠程度不同,有经验的用户或权威用户能够给资源更准确、更客观的评价。可见,用户对资源评价的可靠程度(即用户的权威性),反映了用户评价的稳定性。因此权威用户对资源的评价更值得其他用户参考。

4 结束语

针对用户评分数据稀疏以及系统可扩展性的问题,本文提出了基于分类的协同过滤方法,并通过实验与传统的协同过滤算法进行了有效性的比较分析。实验结果表明,基于分类的协同过滤算法有效地解决了数据稀疏问题,明显地提高了推荐质量,具有一定的有效性和可行性。

参考文献

1 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[J],软件学报,2003,14(9):1621-1624.

2 Shardanand U, Maes P. Social Information Filtering: Algorithms for Automating "Word of Mouth". In: Proceedings of the ACM CHI'95 Conference on Human Factors in Computing Systems, 1995: 210-217.

3 Breese J, Hecherman D, Kadie C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI'98), 1998: 43-52.

4 Ganesan P, Garcia-molina H. Hierarchical Domain Structure to Compute Similarity. ACM Transaction on System, 2003, 21(1): 64-93.

5 Rodriguez M A, Egenhofer M J. Determining Semantic Similarity Among Entity Classes from Different Ontologies. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(2): 442-456.