电子商务协作过滤推荐技术的算法研究与改进

贺智明,王海超,高 娟
(江西理工大学 信息工程学院,江西 赣州 341000)

摘 要:推荐算法的好坏直接影响推荐系统的效率。本文提出了一种改进的基于K-中心点算法的合作聚类推荐算法,该算法有效减少了数值矩阵的行数,大大缩短了搜寻近邻客户的时间,从而提高了算法的执行效率和准确性。

关键词:电子商务;推荐系统;K-中心点算法;客户关系管理

中图分类号:TP301.6 **文献标识码**:A

Research and improvement of e-commerce collaborative filtering recommendation algorithm

HE Zhi Ming, WANG Hai Chao, GAO Juan
(College of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China)

Abstract: The fair or foul of recommendation algorithm can directly affect the recommendation system's efficiency. This aticle propounds an improved cooperation clustering recommendation algorithm based on K-medoids. This algorithm can effectively reduce columns of value matrix, and decrease the time to search near neighbour customers. That increases the algorithm's exeuse effciency and accuracy.

Key words: e-commerce; recommendation system; PAM; CRM

个性化推荐系统是现代商务发展的产物,协作过滤推荐技术是个性化推荐系统中的一种典型技术,其优势是为电子商务的顾客提供个性化服务,促进一对一的销售,使公司拥有顾客的更准确的模型,从而可以对顾客的需求有更好的了解。而服务于这些需求则可在相关产品的交叉销售、提升销售、产品亲和力、一对一促销、保留客户等方面可获得巨大的成功。

然而,协作过滤推荐技术也还存在一些致命的缺点,如稀疏问题、冷开始问题、假负和假正等问题。稀疏问题(Sparsity)是协作过滤推荐技术中的重要问题之一,每个用户一般都只对很少的项目作出评价,整个数据阵变得非常稀疏,一般都在1%以下。这种情况带来的问题是得到用户间的相似性不准确,邻居用户不可靠。冷开始问题又称第一评价问题或新项目问题,如果一个新项目很少有人去评价它,或都不去评价它,则这个项目肯定得不到推荐,推荐系统就失去了作用。假负是指系统没有推荐但顾客却喜欢的产品;假正则是指系统推荐但顾客却并不喜欢的产品[1]。这些问题都不是人们想看到的。因此,怎样使这些问题得到有效的解决就成为目前研究的重点。

1 协作过滤推荐算法

协作过滤推荐算法(Collaborative Filtering Recommendation)是目前应用广泛且效率较高的一种个性化推荐技术。它基于邻居用户的资料得到目标用户的推荐,其推荐的个性化程度更高[2]。

1.1 协作过滤算法的思路

协作过滤推荐是基于邻居用户的兴趣爱好预测目标用户的兴趣偏好。算法首先使用统计技术寻找与目标用户具有相同喜好的邻居,然后根据目标用户的邻居的偏好产生向目标用户的推荐[2]。

协作过滤是基于这样一种假设[3]:如果用户对一些项目的评分比较相似,则他们对其他项目的评分也比较相似;如果大部分用户对一些项目的评分比较相似,则当前用户对这些项目的评分也比较相似。

协作过滤推荐系统使用统计技术搜索目标用户的若干最近邻，然后根据最近邻对项目的评分预测目标用户对项目的评分，产生对应的推荐列表。

1.2 算法模型

对用户已经购买过的商品进行建模，可以有效度量用户之间的相似性。用户评分数据可以用一个 $n\times m$ 阶用户-项目评分矩阵表示，n 行代表个 n 用户，m 列代表 m 个项目，第 i 行 j 列的元素代表用户 i 对项目 j 的评分值。这里只介绍用户间的相似度度量公式，项目间的度量公式和用户间的有些相似。

度量用户间相似性的方法有许多种，主要有 4 种方法：余弦相似性度量公式(Cosine-based Similarity)、修正的余弦相似度公式(Adjusted Cosine Similarity)、相关相似度公式(Correlation-based Similarity)、求熵(互信息)的方法。通常采用前 3 种方法。首先得到用户 i 和用户 j 评分过的所有项目，然后通过不同的相似性度量方法计算用户 i 和用户 j 之间的相似性，记为 $sim(i,j)$。

(1)余弦相似性度量

用户评分看作为 n 维项空间上的向量，如果用户对其项没有进行评分，则将用户对该项的评分设为 0，用户间的相似性通过向量间的余弦夹角度量。设用户 i 和用户 j 在 n 维项空间上的评分分别表示为向量 $\vec{i}$、$\vec{j}$，则用户 i 和用户 j 之间的相似性 $sim(i,j)$ 为：

$$sim(i,j)=\cos(\vec{i},\vec{j})=\frac{\vec{i}\cdot\vec{j}}{\|\vec{i}\|*\|\vec{j}\|}$$

式中，分子为 2 个用户评分向量的内积，分母为 2 个用户向量模的乘积。

(2)修正的余弦相似性

修正余弦相关性充分考虑了不同用户的评分尺度问题，通过减去用户对项目的评分来实现它的优点。设用户 i 和用户 j 评分过的相集合，则用户 i 和用户 j 之间的相似性 $sim(i,j)$ 为：

$$sim(i,j)=\frac{\sum_{c\in I_{i,j}}(R_{i,c}-\vec{R}_i)(R_{J,C}-\vec{R}_J)}{\sqrt{\sum_{C\in I_i}(R_{i,c}-\vec{R}_i)^2}\sqrt{\sum_{C\in I_j}(R_{j,c}-\vec{R}_j)^2}}$$

式中，$R_{i,c}$ 为用户 i 对 c 的评分，$\vec{R}_i$ 为用户 i 的平均评分。

最近邻居查询的目标就是对每一个用户 a，在整个用户空间中查找用户集合，$N=\{N_1,N_2,N_3,\cdots,N_s\}$，$a\notin N$，使得 N_1 与 a 的相似度 $sim(a,N_1)$ 最高，N_2 与 a 的相似度 $sim(a,N_2)$ 次之，依此类推。

(3)相关相似度

相关相似度又称 Pearson 相关系数度量，设用户 i 和用户 j 共同评分过的项目集合用 $I_{i,j}=I_1\cap I_2$ 表示，则用户 i 和用户 j 的相似度 $sim(i,j)$ 为：

$$sim(i,j)=\frac{\sum_{c\in I_{i,j}}(R_{i,c}-\vec{R}_i)(R_{J,C}-\vec{R}_J)}{\sqrt{\sum_{C\in I_i}(R_{i,c}-\vec{R}_i)^2}\sqrt{\sum_{C\in I_j}(R_{j,c}-\vec{R}_j)^2}}$$

1.3 邻居集合的形成

邻居集合的形成一般有 4 种方法：$Top-N$、K 近邻法、阈值法、聚类法、贝叶斯网络法。最常用的是前 2 种方法。

算法的核心部分是为一个需要推荐服务的目标用户寻找最相似的最近邻居集。根据预先确定的邻居数 N，采用以上相似度的算法按由大到小的顺序选取前 N 个用户作为邻居用户集合。或者根据预先确定的相似度阈值，选择所有相似度大于阈值的作为邻居用户集合。

1.4 推荐产生

根据当前用户最近邻居对商品的评分信息预测当前用户对未评分商品的评分，产生 $Top-N$ 商品推荐。通过上面提出的相似性度量方法得到目标用户的最近邻居，下一步需要产生相应的推荐。设用户 u 的最近邻居集合用 N_u 表示，则用户 u 对项目 i 预测评分 $P_{u,i}$ 可以通过用户 u 对最近邻居集合 N_u 中项的评分得到，计算方法如下：

$$P_{u,i}=\bar{R}_u+\frac{\sum_{n\in NN_u}sim(u,n)\times(R_{n,i}-\bar{R}_n)}{\sum_{n\in NN_u}(|sim(u,n)|)}$$

式中，$sim(u,n)$ 表示用户 u 与用户 n 之间的相似性，$R_{n,i}$ 表示用户 n 对项 i 的评分，R_u 和 R_n 分别表示用户 u 和用户 n 对项的平均评分。

通过上述方法预测用户对所有未评分项的评分，然后选择预测评分最高的前 n 项作为推荐结果反馈给当前的目标用户。

2 基于 K-中心点算法的合作聚类算法

尽管协作过滤技术在个性化推荐系统中获得了极大的成功，但随着电子商务系统规模的扩大，用户数目和项数目指数级增长，导致用户评分数据的极端稀疏性。由于用户的最近邻居至少对 2 件商品进行了共同评分，因此在用户评分数据极端稀疏的情况下，无法搜索到某些用户其最近邻居，导致协作过滤推荐算法无法对这些用户产生任何推荐。其次，在大规模数据集上搜索当前用户的最近邻居非常费时，难以保证协作过滤推荐算法的实时性要求。最后，协作过滤推荐算法无法发现商品之间存在的隐含关联[4]。

现有许多种改进的算法来解决这一难题，如基于降维的协作过滤推荐算法、Cluster-based 协作过滤推荐算法都是目前的主流算法。在基于降维的协作过滤推荐算法中，奇异值分解 SVD(Singular Value Decomposetion)技

术在信息检索领域得到了广泛应用。基于SVD技术的协作过滤推荐算法能较好地解决数据稀疏性问题，同时，因为$k<<n$，计算开销也相应降低了，这有利于解决推荐算法的伸缩能力问题，但推荐的精确性也会因此有所下降。

2.1 Cluster-based协作过滤推荐算法

Cluster-based协作过滤推荐算法，将整个Web日志根据用户的购买习惯和评分特点划分为若干个不同的聚类，从而使得聚类内部用户对项的评分尽可能相似，而不同聚类间用户对商品的评分尽可能不同甚至相反。使目标用户与其相似度最近的那个簇对其进行推荐，从而提高了精确度，也提高了最近邻查询的效率。

根据每个聚类中用户对商品的评分信息生成一个虚拟用户，它代表了该聚类中用户对商品的典型评分，将所有虚拟用户对商品的评分作为新的搜索空间，查询当前用户在虚拟用户空间中的最近邻居，产生对应的推荐结果。相对于原始的用户空间而言，虚拟用户空间要小得多，因此最近邻查询的效率也高得多，可以有效提高推荐算法的实时响应速度[4]。

2.2 改进的基于K-中心点算法的合作聚类算法

本文提出了一种改进的K-中心点算法(PAM)用来对整个用户的访问记录和访问特点进行聚类，主要步骤如下：

设站点有m个页面，共有n个用户访问，由于采用协作推荐方法，设T为一个$n\times(m+1)$的矩阵。$n\times m$的矩阵为用户-项目矩阵。第$m+1$列表征该行被加入到该矩阵中的时间，目的是为了始终让此矩阵保持最新状态，避免一些过时的兴趣，因为客户的兴趣可能会改变。

输入：初始簇K、T。

输出：生成新的聚类中心Maincenter。

(1)$k=[K/2]$; //起始时取$[K/2]$值作为k-中心点算法的初始k值

(2)随机选取k个对象作为初始的簇的中心。

(3)重复。

(4)对其他非中心点对象，计算其与中心点的距离，并将其分配到距离最近的中心点代表的簇。

(5)重复。

(6)选择一个未被选择的中心点O_i。

(7)重复。

(8)选取一个未被选择的非中心点对象O_m，计算用O_m代替O_i的总代价并记录在集合S中。

(9)直到所有的非中心点对象都被选择过。

(10)直到所有的中心点都被选择过。

(11)若在S集合中所有非中心点对象代替所有中心点后计算的总代价中存在小于0的，则找出S中最小的一个，用该非中心点替代对应的中心点。

(12)若在S集合中所有非中心点对象代替所有中心点后计算的总代价中存在大于0的，则找出代价最大的一个，并将其设为一个新的中心点。

(13)这样形成一个新的含有$k+1$个中心的集合。

(14)直到S集合中所有的值都大于0，且$k<=K$。

(15)最后将每个用户分配到相似性最高的聚类中。

(16)对新生成的聚类，计算聚类中所有用户对项的平均评分，生成新的聚类中心。

(17)重复15~16，直到聚类不再发生改变为止。

生成聚类之后，Cluster-based协作过滤推荐算法可以分为如下2步：

(1)生成虚拟用户集

虚拟用户集由聚类所得的聚类中心组成，这些聚类中心是根据不同的聚类生成的，是每个聚类中与其他用户的距离之和最小的对象的集合，代表了其所在聚类中用户对商品的典型评分。

(2)产生推荐

得到虚拟用户集之后，对其使用各种相似性度量方法以搜索当前用户的最近邻居，再根据这些最近邻居对商品的评分信息来生成相应的推荐结果。其方法与协作过滤推荐算法类似，不再赘述。

由于采用了聚类算法压缩了T矩阵(减少了行的个数)，当一段时间之后，一些新的用户访问被换入T矩阵后，就需要重新运行此算法已得到新的压缩结果。

电子商务已经成为现代商务的主流，其规模也已变得越来越大，伴随着商品同质化时代的来临，提高客户的满意度、忠诚度，将是企业盈利的首要因素，对于推荐系统的要求也将越来越高。本文通过将K-中心点算法与合作聚类算法融合，可有效解决传统推荐系统中的冷开始、数据稀疏性、假负、假正等问题，从而可以更好地获得相近客户，提高推荐的效果和准确性。

参考文献

[1] HAN J W, KAMBER M. 数据挖掘概念与技术[M]. 范明，孟小峰，译. 北京：机械工业出版社，2007：440.

[2] 鲁为. 协作过滤算法及其在个性化推荐系统中的应用[D]. 北京：北京邮电大学，2007:22-24.

[3] BREESE J, HECKERMAN C. kadie. Empirical analysis of predictive allgorithms for collaborative filtering. In: Proceedings of the 14th Conference on Uncertinty in Aritificial Intelligence, San Francisco, CA, July 1998: 44-52.

[4] 邓爱林. 电子商务推荐系统关健技术研究[D]. 上海：复旦大学，2003.

(收稿日期：2009-03-03)