

# 电子商务推荐系统中的协同过滤推荐

游文, 叶水生

(南昌航空工业学院 计算机科学与技术学院, 江西 南昌 330034)

**摘要:** 电子商务推荐系统中协同过滤已成为目前应用最广泛、最成功的推荐方法。它利用相似用户购买行为也可能相似的特性进行推荐。介绍了与其他方法比较协同过滤方法的优点, 然后说明了一些主要的协同过滤实现方法, 指出了还需改进和完善的地方以及未来研究的方向。

**关键词:** 电子商务; 推荐系统; 协同过滤

**中图分类号:** F724.6

**文献标识码:** A

**文章编号:** 1673-629X(2006)09-0070-03

## A Survey of Collaborative Filtering Algorithm Applied in E-commerce Recommender System

YOU Wen, YE Shui-sheng

(Department of Computer, Nanchang Institute of Aeronautical Technology, Nanchang 330034, China)

**Abstract:** In E-commerce recommender system, collaborative filtering technology is the most popular and successful method at present. It supposes similar users may have the same behavior in shopping. In this article, first introduce strong-points of the algorithms comparing with other methods, then describe several main collaborative filtering algorithms, at last, point out several open research problems and directions on the algorithm.

**Key words:** E-commerce; recommender system; collaborative filtering algorithm

### 0 引言

电子商务对传统的商务交易产生了革命性的变化, 从而要求“以产品为中心”向“面向客户”、“以客户为中心”的新的商业模式的转变, 要求电子商务网站按客户群划分产品, 围绕客户进行服务, 为客户提供所需要的东西, 所以对每个顾客提供个性化的服务成为必要。在这种背景下, 推荐系统(Recommender System)应运而生, 它是根据用户个人的喜好、习惯来向其推荐信息、商品的程序<sup>[1]</sup>。电子商务网站可以使用推荐系统分析客户的消费偏好, 向每个客户具有针对性地推荐产品, 帮助用户从庞大的商品目录中挑选真正适合自己需要的商品, 尽可能为每个顾客提供个性化的服务。

目前推荐系统中使用的主要推荐技术有协同过滤推荐、基于内容推荐、基于人口统计信息推荐、基于知识推荐和基于规则推荐等。协同过滤推荐(collaborative filtering recommendation)是目前研究最多、应用最广的个性化推荐技术。

### 1 协同过滤算法

协同过滤的出发点是<sup>[2,3]</sup>: 兴趣相近的用户可能会对同样的东西感兴趣。所以, 只要维护关于用户喜好的数据, 从中分析得出具有相似口味的用户, 然后就可以根据相似客户的意见来向其进行推荐。另一种可能的出发点是: 用户可能较偏爱与其已购买的东西相类似的商品。可以根据用户对各种东西的评价来判断商品之间的相似程度, 然后推荐与用户兴趣最接近的那些商品。前一种思路以客户与客户之间的关系为中心, 而后一种思路则以项目与项目之间的关系为着眼点。协同过滤推荐的个性化程度高, 目前有许多网站采用了基于该技术的推荐系统, 如 Amazon.com, CDNow.com, MovieRinder.com 等。

#### 1.1 协同过滤的优点

与其他方法相比, 协同过滤具有下列优点<sup>[4]</sup>:

- (1) 能跨类型推荐, 如艺术品、音乐、电影等。
- (2) 不需要领域知识, 共享其他人的经验。
- (3) 自适应性好: 随时间推移, 推荐质量会提高。

(4) 充分的隐式反馈, 能够减少用户的反馈量, 加快个性化学习的速度。

#### 1.2 协同过滤的输入与输出

在协同过滤中需要所有用户对各种产品的看法作为判断的依据。假设有  $m$  个用户  $\{u_1, u_2, \dots, u_m\}$  和  $n$  个项目  $\{i_1, i_2, \dots, i_n\}$ , 原始数据通常表示为  $m \times n$  的二维矩

收稿日期: 2005-12-20

**作者简介:** 游文(1977-), 男, 江西景德镇人, 硕士研究生, 研究方向为 Web 数据挖掘、人工智能; 叶水生, 硕士, 教授, 研究方向为 Web 数据挖掘、人工智能。

阵,矩阵中的值  $v_{ij}$  是用户  $i$  对项目  $j$  所作的评分,用某一范围的整数表示,如  $1 \sim 5$ , 0 则表示用户尚未对项目作出评价。获取这些数据有各种途径,可以像 GroupLens 和 Ringo 系统所做的那样,直接要求用户对所看见的商品进行打分,也可以从用户的历史购买记录、收藏夹等处隐性提取。

把推荐算法当前推荐的对象称为目标客户,记作  $u_n$ , 推荐算法的结果可以用两种形式来表示:预测目标客户  $u_n$  对项目  $j$  的评分,结果是个数值;或是目标客户最感兴趣的  $N$  个产品的推荐列表,当然这  $N$  个产品必须是目标客户还没有购买过的<sup>[3]</sup>。

### 1.3 协同过滤算法的分类

近些年来有许多学者提出了各种协同过滤的实现算法,如果按照协同过滤算法出发点的不同,则可分为基于客户-客户关系的协同过滤算法和基于项目-项目关系的协作过滤算法。

(1) 基于客户-客户关系的协同过滤算法<sup>[5]</sup>。

“最近邻居算法”是到目前为止最为成功的自动推荐技术,被许多系统采用。这种技术使用统计方法挑选出与目标用户最相似的若干用户,称为“邻居”,然后根据这些邻居的意见推测用户对目标商品感兴趣程度。

那么如何估算用户之间的相似性呢?一种方法是计算相关系数。以 Pearson 相关系数为例。用户  $a$  和用户  $b$  之间的相关系数的计算公式如下:

$$\text{corr}_{ab} = \frac{\sum_{j \in I_{a,b}} (v_{a,j} - \bar{v}_a)(v_{b,j} - \bar{v}_b)}{\sqrt{\sum_{j \in I_{a,b}} (v_{a,j} - \bar{v}_a)^2} \sqrt{\sum_{j \in I_{a,b}} (v_{b,j} - \bar{v}_b)^2}}$$

其中  $v_{ij}$  代表用户  $i$  对项目  $j$  所作的评分,  $\bar{v}_a, \bar{v}_b$  分别是用户  $a$  和用户  $b$  所有打过分的项目的平均得分。另一种方法是将关于用户  $a$  和用户  $b$  的记录看作是两个向量,则可以用向量之间夹角的余弦值来表示用户的相似度。

$$\cos(a, b) = \frac{a \cdot b}{\|a\|_2 * \|b\|_2}$$

有了相似度的衡量标准,下面就该确定目标用户的  $l$  个邻居了。这里也有两种方法可供选择。

a. 直接以目标用户为中心,找出与之最相似的  $l$  个用户。

b. 用聚集的方式,首先找出离目标用户最近的用户,然后依次找出余下的用户。假设已经找到  $j$  个邻居,计算这  $j$  个邻居的中心位置  $\bar{C} = \frac{1}{j} \sum_{i=1}^j v_i$ , 选择其他用户中与该中心位置最近的用户作为第  $j+1$  个邻居。如此反复,直到  $j = l$  为止。这种方式较适合于稀疏数据集。

最后计算各邻居对商品  $j$  评分的加权和,权值大小根据相关程度来确定,这个值就可作为目标客户对项目  $j$  评分的预测值。具体的计算公式如下:

$$P_{aj} = \bar{v}_a + k \sum_{i=1}^l \text{corr}_{ai} (v_{ij} - \bar{v}_i)$$

式中,  $k$  是用来规范化权值的因子。

(2) 基于项目-项目关系的协同过滤算法<sup>[3]</sup>。

与基于客户-客户关系的协同过滤算法不同,基于项目-项目关系的协同过滤算法首先关注的是项目之间的联系。算法查看所有目标客户已经评价的项目集合,计算这些项目与正考虑推荐的项目  $i$  之间的相似程度,并从中挑选出最相似的  $k$  个项目  $\{i_1, i_2, \dots, i_k\}$ , 其对应的相似程度为  $\{s_{i_1}, s_{i_2}, \dots, s_{i_k}\}$ 。然后计算目标客户对这些相似项目的评分的加权平均值,即得到所需的预测值。算法的实现可以分为相似度计算和预测两个阶段。

第一步相似度的计算完全类似于基于用户之间关系的最近邻居算法,不同之处在于最近邻居算法计算矩阵的行之间的相似度,而基于项目之间关系的算法计算列与列之间的相似度。具体地,如果现在要计算项目  $i$  与项目  $j$  之间的相似度,首先要做的是找出所有同时评估这两个项目的用户,得到关于这两个项目的两个列向量,然后选择相关系统或余弦值计算两个向量之间的相似度,公式与前类似,不再赘述。由于余弦值的方法中没有把各个用户打分的松紧差异考虑进去,补偿的方法是从每对值中减去用户的平均值,计算公式为:

$$s(i, j) = \frac{\sum_{u \in U} (v_{ui} - \bar{v}_u)(v_{uj} - \bar{v}_u)}{\sqrt{\sum_{u \in U} (v_{ui} - \bar{v}_u)^2} \sqrt{\sum_{u \in U} (v_{uj} - \bar{v}_u)^2}}$$

式中  $U$  是所有同时标项目  $i$  和项目  $j$  的用户集合。

分别计算出目标客户所有已经评价的项目与要预测的项目  $i$  之间的相似度后,过滤出前  $k$  个值最大的项目  $N$ , 然后计算这些值的和,每个值以相似度为权。具体公式如下:

$$P_{ai} = \frac{\sum_{j \in N} (s_{ij} \cdot v_{aj})}{\sum_{j \in N} (|s_{ij}|)}$$

## 2 待解决的问题

实验表明协同过滤算法可以提供较准确的推荐,但是在运用中也暴露出一些需要解决的问题<sup>[1,5]</sup>。

### 2.1 数据稀疏性问题

电子商务网站通常拥有大量商品,而每个用户购买或作评价的只是其中很小的一部分,通常不到 1%,数据的稀疏会导致算法准确率降低。而相关系数需要两个用户至少评价了两个以上相同的产品才能计算,两个实际上相似的用户很可能因为缺乏相同的产品而失之交臂,算法甚至找不到任何商品可作推荐。

目前的一种解决途径是在运用具体的协同过滤算法之前先使用维归约技术对原始数据进行压缩,在降低维数的同时使信息损失降为最小。维归约技术一方面可以提高过滤算法的效率,另一方面通过充分发掘隐含在数据中的潜在模式,突破协同过滤算法需要相关用户标注相同的项目的限制。可以考虑的维归约方法有 SVD (Singular Value Decomposition) 奇异值分解、基于信息增益的属性相

关分析<sup>[6]</sup>等。

## 2.2 扩展性问题

电子商务中的运用涉及的数据量非常庞大,而像最近邻居这样的算法效率随着用户和产品的数目增多而下降,如何改进算法使之更适应大规模计算是研究的重点之一。解决这个问题常采用的是用户筛选技术,一方面可以通过给不同推荐赋予不同的权重,另一方面可以在最近邻居用户的选择上作一定的改进。具体实现方法一般有两种方法:一种是选取具有新颖描述(Novel Profile)的用户,另一种是选取具有合理描述(Rational Profile)的用户。

## 2.3 实时性

实时性问题随着需要在线服务的客户大量增加而显得越来越突出,解决这个问题的有效方法是采用分布式计算技术。另外基于项目聚类的协同过滤推荐算法<sup>[7]</sup>可以显著缩小最近邻居的查询空间,从而有效解决推荐系统处理大规模数据面临的实时性问题。

## 2.4 评分数据不足

过多地要求用户主动提供数据会使用户感到不便而转向其他网站,可以借助于半智能的代理程序隐式地记录用户的行为,分析提取喜爱的模式。数据挖掘中的 Web 使用记录挖掘是很好的辅助工具。另一种解决途径是采用基于项目评分预测<sup>[4]</sup>的方法对未评分项目进行预测。

除了上述问题之外,推荐系统还有许多其他可以改进的地方。协同过滤方法只考虑客户、项目之间的关系,丝毫不考虑项目本身的特点。现实中可能会影响顾客决定的因素很多,只有充分地理解用户才能更贴近用户的真实需要,推荐算法应尽可能地利用各种有用的资料信息,例如把基于内容的筛选与协同过滤结合在一起,互取所长。

以客户为中心,可以实现更灵活的推荐形式:如果一个用户想给朋友买一件礼物,推荐系统可以根据朋友所属的类别作推荐;多个人想一起看电影,这时推荐的影片必须至少不被任何一个人讨厌,这是针对客户群的推荐;如

果在推荐的同时说明推荐的理由,也许会更具有说服力.....<sup>[1]</sup>

## 3 小 结

电子商务领域中,推荐系统在帮助用户快速定位感兴趣的商品的同时也为企业实现了增值,已经成为电子商务网站的关键模块。而协同过滤与其他推荐方法比较有许多不可替代的优点,文中对协同过滤算法进行了简单介绍,并对存在的问题进行了概述。目前国内的电子商务网站在这方面的实践处在快速发展的阶段,因此还需要继续研究其更智能、更优化的协同过滤模型及算法。

## 参考文献:

- [1] Ben S J, Konstan J A, Riedl J. E-commerce Recommendation Applications[EB/OL]. <http://www.grouplens.org/papers/pdf/ECRA.pdf>, 2002-09-26.
- [2] Breese J S, Heckerman D, Kadie C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering[A]. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence[C]. Madison, Wisconsin: Morgan Kaufmann, 1998. 43-52.
- [3] Badrul S, Karypis G, Konstan J, et al. Item-based Collaborative Filtering Recommendation Algorithms[Z]. WWW10, 2001.
- [4] 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): 1621-1628.
- [5] Badrul S, Karypis G, Konstan J, et al. Analysis of Recommendation Algorithms for E-commerce[M]. New York, USA: ACM Press, 2000. 106-112.
- [6] 赵 亮,胡乃静,张守志. 个性化推荐算法设计[J]. 计算机研究与发展, 2002, 39(8): 986-991.
- [7] 邓爱林,左子叶,朱扬勇. 基于项目聚类的协同过滤推荐算法[J]. 小型计算机系统, 2004, 25(9): 1665-1670.

(上接第 69 页)

## 5 结束语

针对不完备决策信息系统,给出在容差关系下属性相对约简的定义,提出了求取不完备决策表系统属性相对约简算法。并通过实例说明该算法能找到决策表的相对最小约简。关于最小约简的完备性还有待于进一步探讨。

## 参考文献:

- [1] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Science, 1982, 11(5): 341-356.
- [2] 束志恒,陈德钊,陈亚秋. 粗糙集方法及其在化学模式分类规则挖掘中的应用[J]. 分析化学, 2004, 32(7): 879-883.
- [3] 钟 波,周家启,肖 智. 基于粗集与神经网络的电力负荷新型预测模型[J]. 系统工程理论与实践, 2004, 24(6): 113-119.

- [4] Pawlak Z, Grzymala B J, Slowinski R, et al. Rough sets[J]. Communication of the ACM, 1995, 38(11): 89-95.
- [5] 刘少辉,盛秋骛,吴 斌,等. Rough 集高效算法研究[J]. 计算机学报, 2003, 26(5): 524-529.
- [6] 王国胤,于 洪,杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7): 759-766.
- [7] 黄 兵,周献中,张蓉蓉. 基于信息量的不完备信息系统属性约简[J]. 系统工程理论与实践, 2005, 25(4): 55-60.
- [8] 王国胤. Rough 集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001.
- [9] 徐德友,胡寿松. 一种基于粗糙集的近似质量求取属性约简的决策算法[J]. 控制与决策, 2003, 18(3): 313-316.
- [10] KRYSZKIEWICZ M. Rough set Approach to Incomplete Information systems[J]. Information Sciences, 1998(112): 39-49.