基于协同过滤的推荐系统算法研究

刘景昊
华中科技大学文华学院 湖北 武汉 430074

【摘 要】在网络购物日益风靡的今天，怎样提供优质的个性化服务是当今电子商务系统的核心内容，而协同过滤推荐则是当今发展最成熟且最成功的推荐系统。本文将全方位介绍协同过滤推荐的内容、研究成果以及协同过滤算法中出现的问题，并提出协同过滤算法进一步发展的方向。

【关键词】协同过滤 推荐系统 发展方向
中图分类号：TP301.6 文献标识码：A 文章编号：1009-4067(2013)06-157-01

1、引言

随着网络技术的发展，人们正处于一个信息爆炸的时代。协同推荐系统（或协同过滤系统）尝试基于其他用户的过往使用情况来预测某个产品对于某个特定的用户的效用。大部分通俗的做法，都是某个产品 s 对于某个用户 c 的效用 $u(c_j,s)$ 的评价标准都是基于产品 s 对其他一些“类似”用户 c 的用户 c_j 的效用 $u(c_j,s)$，如在一个电影推荐程序中，为了给用户 c 推荐电影，协同推荐系统首先尝试找出用户 c 的同类群，也就是其他有类似欣赏风格的用户，然后，把只有最相近于用户 c 的同类群的电影才会被推荐。

被认为最早的推荐系统是 Grundy 系统[1]。接下来的 Tapestry 系统依靠每名用户的用户手册辨认志趣相投用户。Group Lens[2]，Video Recommender [3] 和 Ringo [4]是最早使用协同过滤算法来进行自动预测的，其它协同系统包括 Amazon.com 的书籍推荐，帮助人们在万维网上搜索相关信息的 Phoaks 以及推荐笑话的 Jester。

2、算法介绍

协同过滤算法可以分为两大类：以记忆行为为基础与以模型为基础。基于记忆行为的算法 [5]，其[6]本质上是基于用户先前相关条目的完整集合作一个关联度的预测。对于用户 c 已经产品 s 的未知关联度数值 $r_{c,s}$ 是其他用户对于同一样产品 s 的关联度的总和。

$$r_{c,s} = aggr_{c'\in\hat{C}'} r_{c',s}, \quad (1)$$

$\hat{C}$ 表示集合 N 是对于关联性产品 s 与用户 c 有最大相似性的用户群。下面是一些计算公式的例子：

$$\begin{aligned}&\text{(a) } r_{c,s} = \frac{1}{N}\sum_{c'\neq c} r_{c',s},\\ &\text{(b) } r_{c,s} = k\sum sim(c,c')\times r_{c',s},\\ &\text{(c) } r_{c,s} = \bar{r}_c + k\sum sim(c,c')\times r_{c',s},(r_{c',s}-\bar{r}_{c'}).\end{aligned} \quad (2)$$

乘数 k 在这里是提供一个常数项，通常是用公式：$k = 1/\sum_{c'\in\hat{C}} sim(c,c')$ 表示。而对于用户 c 的平均关联度，在（10c）中是定义为：

$$\bar{r}_c = (1/|S_c|)\sum_{s\in S_c} r_{c,s}, where S_c = \{s\in S | r_{c,s}\neq\varnothing\}. \quad (3)$$

在这种情况中是用与平均值的相差量的权重总和来代替那些完全不变的评级。另一种克服使用不同评级量表的方法是：调整基于优先偏好的过滤，也就是专注于推荐用户相关优先来替代绝对等级值。

3、相似度计算

在协同推荐系统中，有很多方法被用来计算用户之间的相似度。在协同推荐系统中，S_w 通常作为是计算用户 x 的“近邻”的中间结果，以及用一种直接的方式来进行计算，例如作为计算集合 S_x 和集合 S_y 的交叉数据。然而，对于某种计算方法，如把 graph-theoretic（图论）用于协同过滤可以在无需计算所有 y 用户的 S_w 的情况得到关于 x 的近邻值。在基于相关性的方法中，Pearson 相关性的有效程度通常被用来衡量他们之间的相似度[7]：

$$sim(x,y) = \frac{\sum_{s\in S_w}(r_{x,s}-\bar{r}_x)(r_{y,s}-\bar{r}_y)}{\sqrt{\sum_{s\in S_w}(r_{x,s}-\bar{r}_x)^2\sum_{s\in S_w}(r_{y,s}-\bar{r}_y)^2}}. \quad (4)$$

在基于余弦值计算的方法中，用户群 x 与 y 被看作是在 m 维空间中的两个向量。而对于这两个向量的相似度，就可以通过计算它们之间的斜率的余弦值来得到：

$$sim(x,y) = \cos(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{\|\vec{x}\|_2\times\|\vec{y}\|_2} = \frac{\sum_{s\in S_w} r_{x,s}r_{y,s}}{\sqrt{\sum_{s\in S_w} r^2_{x,s}}\sqrt{\sum_{s\in S_w}(r^2_{y,s})}} \quad (5)$$

且 $\vec{x}\cdot\vec{y}$ 表示在向量 $\vec{x}$ 与 $\vec{y}$ 之间的所有点所对应的产品。还有其他的处理方法就是通过度量用户群之间的平均方差值来计算相似度。不同的推荐系统有不同的方法来提升在用户相似度计算以及评级估算的有效性。一个很普遍的策略就是预先（因为网络群体通常不会在短时间内发生剧烈改变）计算所有用户的总体相似度 $sim(x,y)$（包括计算 S_w），然后再分别重新计算每个用户的相似度。因此，无论用户何时对系统发出请求，它的评级都能通过实时按需计算而有效的得到。一种不同的方法被采用，目的是来改善已经存在的协同过滤算法的性能，在这个方法中用户组评级的输入是被精心挑选的，使用的技术包括排除噪音、冗余度还有就是利用了评级数据的稀疏性。此外，在其中的最新发展中，提出了一种协同过滤的概率方法，即把基于记忆与基于模型的技术结合起来的方法。

4、存在问题

1 新用户问题。这是和基于内容的推荐系统共同存在的问题。为了做精确的推荐，系统首先必须从用户给出的评级了解用户的偏好。几种技术已经被提出来了去解决这个问题。其中的大部分都使用混合推荐方法即把基于内容和协同的技术结合起来。另一种方法被描述，为了决定给新用户推荐最好的项目而探讨了各种不同的技术。这些技术使用基于项目流行度、项目信息熵、用户个性化的策略，并且组合了上述方法。

2 新项目问题。新项目被频繁地增加到推荐系统中。协作系统单纯地依靠用户的偏好去做推荐。因此，直到新的项目被大量的用户评级后推荐系统才能推荐它。这个问题也能通过使用混合推荐方法被解决。

3 稀疏性。在任何推荐系统中，已经获得的评级数目通常相比于需要预测的评级数目是很小的。从较小数量的事例中获得有效的等级预测是很重要的。协同推荐系统的成功还要依赖足够数量用户的可得性。一种克服评级稀疏性问题的方法就是使用用户文件信息去计算用户的相似度。也就是说，二个用户可以被认为相似，不仅只是当他们把电影评定成了相同的等级时，还有就是当他们属于相同的人群结构段时也可以被认为相似。这种传统的协同过滤拓展有时被称为“人口过滤”。另一种方法也利用了已经在中提过的用户之间的相似度，这里稀疏性问题通过运用关联检索框架和相关的激活扩算算法来解决，目的是通过他们过去的交易和反馈来探索消费者中的可传递的协会。一个不同的处理稀疏评级矩阵的方法被用在中，这里一个降维技术—奇异值分解被用来降低稀疏评级矩阵的维数。

5、总结与展望

本文介绍了协同过滤推荐算法的主要思想及算法需要解决的主要问题，总结了近期本领域内的研究进展。我们从以下几方面对算法的发展方向进行研究：考虑在保证推荐准确性的同时增加推荐的多样性；使特殊喜好的用户得到推荐需要考虑用户对系统的实时反馈，更好的利用反馈数据强化推荐算法；考虑时间，空间，任务等因素，来更好的完成推荐。

参考文献

[1]E. Rich, “User Modeling via Stereotypes,” Cognitive Science, vol. 3, no. 4, pp. 329-354, 1979.

[2]J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl, “GroupLens: Applying Collaborative Filtering to Usenet News,” Comm. ACM, vol. 40, no. 3, pp. 77-87, 1997.

[3]W. Hill, L. Stead, M. Rosenstein, and G. Furnas, “Recommending and Evaluating Choices in a Virtual Community of Use,” Proc. Conf. Human Factors in Computing Systems, 1995.

作者简介

刘景昊（1991-），男，内蒙集宁人，华中科技大学文华学院通信工程专业，研究方向：现代通信网络.

基于协同过滤的推荐系统算法研究

作者：刘景昊
作者单位：华中科技大学文华学院 湖北武汉430074
刊名：中国电子商务
英文刊名：Discovering Value
年，卷(期)：2013(6)

参考文献(3条)

1. E.Rich ″ User Modeling via Stereotypes 1979(04)

2. J.A.Konstan.B.N.Miller.D.Maltz.J.L.Herlocker,L.R.Gordon,and J.Riedl ″GroupLens:Applying Collaborative Filtering to Usenet News,″ 1997(03)

3. W.Hill.L.Stead.M.Rosenstein.G.Furnas ″Recommending and Evaluating Choices in a Virtual Community of Use 1995