**2.3.2　数据稀疏和冷启动问题**

2013-06-23 22:32 蒋凡 译 人民邮电出版社 字号：**[T](javascript:setfont(12);" \t "_self)** | **[T](javascript:setfont(16);" \t "_self)**

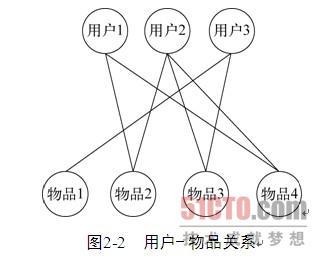
**2.3.2　数据稀疏和冷启动问题**

在前面例子用到的评分矩阵中，只有一个用户 物品组合没有评分。但在实际应用中，由于用户一般只会评价（或购买）少部分物品，评分矩阵一般都非常稀疏。

这种情况下的挑战是用相对较少的有效评分得到准确的预测。直接做法就是利用用户的附加信息，比如性别、年龄、教育程度、兴趣等能够帮助分类用户的信息。因此，相似用户（近邻）集合不只是根据显式或隐式评分，也会根据评分矩阵的外部信息来分析。这些系统（Pazzani(1999b)提到的混合系统）利用了人口统计信息以及已经不再"纯粹"的协同方法。这种方法也引出了新问题，比如如何获取额外信息以及如何混合不同的分类器。尽管如此，在刚上线推荐服务的扩张阶段，这种技术对于获取协同方法所需的大量关键用户还是有帮助的。

多年以来，人们提出过一些处理冷启动和数据稀疏问题的方法。我们在这里要详细讨论的例子是Huang et al.(2004)提出的基于图的方法。其主要思想是利用假定用户品味的"传递性"，并由此增强额外信息矩阵。

考虑图2-2中的用户-物品关系，它由表2-3的二进制评分矩阵（摘自Huang et al.(2004)）推导出。

[](http://s1.51cto.com/wyfs01/M00/0F/87/wKioJlHHB1KS8w90AAA0XWhIE7E626.jpg)

在矩阵中，0不是一个显式的（差评）评分，而是一个缺失的评分。假设我们正在为用户1寻找推荐。根据标准的CF方法，用户2会被认为是用户1的同伴，因为他们都买了物品2和物品4。物品3会被推荐给用户1，因为最近邻用户2也买了或喜欢它。Huang et al.(2004)将这种推荐问题看做图分析问题，推荐由用户和物品间的路径决定。在标准的基于用户或基于物品的CF方法中会考虑长度为3的路径，即物品3和用户1相关是因为他们之间存在一条3步路径（用户1-物品2-用户2-物品3）。由于这种长度为3的路径数量在稀疏评分数据中很少，因此这种思路也会考虑更长的路径（间接关联）计算推荐。比如，由于存在两条连接用户1和物品1的5步路径，利用长度为5的路径也可以推荐物品1。

表2-3　扩展激活方法的评分数据库

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 物品1 | 物品2 | 物品3 | 物品4 |
| 用户1 | 0 | 1 | 0 | 1 |
| 用户2 | 0 | 1 | 1 | 1 |
| 用户3 | 1 | 0 | 1 | 0 |

由于这些远距离关系的计算代价很高，Huang et al.(2004)提出将评分矩阵转化为用户和物品的双向图，并使用一种称为扩展激活的特殊图搜索方法高效地分析图。对比标准的基于用户和基于物品的算法显示，基于间接关系的技术能显著地提高推荐质量，尤其是在评分矩阵稀疏的时候。同样，相比标准的协同过滤技术，该算法对新用户的推荐也能有明显的提升。然而当评分矩阵达到某种密度之后，相比标准算法的推荐质量也会有所下降。尽管如此，距离关系的计算代价仍然很高，现在还无法实际应用在大规模评分数据上。

缺省投票是Breese et al.(1998)描述的另外一种用来处理稀疏评分数据的技术。回想一下，标准的相似度方法只考虑那些当前用户和用来比较的用户都评过分的物品。当数量很少时，评分碰巧相同或不同都会对相似度计算影响很大。因此这种思路就是给那些只有一两个用户评过分的物品赋以缺省值（可能也会对一些附加的物品），这样可以提高稀疏评分数据上的预测质量（Breese et al.1998）。这些人工缺省投票就像一种缓冲机制，能够减少那些个别巧合因素对相似度的影响。

近来，Wang et al.(2006)还提出过其他解决数据稀疏问题的方法。由于发现大多数协同推荐只用到评分数据中的某部分特定信息（用户相似度或物品相似度），因此研究人员建议将这两种不同类型的相似度组合起来提高预测准确率。此外，他们还在预测函数里利用了先前方法没有考虑到的第三种信息，即相似用户给出的相似的物品评分。通过一种概率机制将来自不同数据源的不同预测进行"融合"和平滑，其中第一组实验显示预测准确率得到提高，尤其是在处理稀疏评分数据时。

冷启动问题是稀疏问题的一个特例（Huang et al.2004）。此类问题包括：(1) 如何向还没给任何物品评分的新用户推荐；(2) 如何处理从未被评过分或购买过的物品。这两类问题都可以通过混合方法来解决，即利用额外的外部信息（Adomavicius and Tuzhilin 2005）。对于新用户问题，其他策略也可能奏效。一种方法是在推荐之前要求用户给出最低限度数量的评分。在这种情况下，系统需要能够从信息论角度智能地获取具有最多信息量的物品评分（Rashid et al.2002）。Goldberg et al.(2001)提出的Eigentaste算法要求用户提供标准集合的评分也是一种类似的策略。