基于兴趣度向量模型的协同过滤推荐技术研究

肖满生[1]，王　宏[2]

（1.湖南工业大学 理学院，湖南 株洲 412008；
2.湖南工业大学 包装设计技术专业中心，湖南 株洲 412008）

摘　要：在引入兴趣度向量模型的基本原理和用户协同过滤推荐技术的基础上，探讨一种有效的基于兴趣度向量模型的商品推荐算法。该算法通过将兴趣度向量模型与协同过滤推荐技术结合起来，在电子商务中实现商品自动推荐，从而提高推荐精度和推荐质量。并对这种推荐算法的有效性进行了实验验证。

关键词：兴趣度向量模型；协同过滤；商品推荐

中图分类号：TP311.51　　**文献标识码**：A　　**文章编号**：1673-9833(2008)04-0041-03

Research on Coordination Filtration Recommendation Technology Based on Interest Vector Model

Xiao Mansheng[1]，Wang Hong[2]

（1. School of science，Hunan University of Technology，Zhuzhou Hunan 412008，China；
2. The Center of Packaging Design Specialty，Hunan University of Technology，Zhuzhou Hunan 412008，China）

Abstract: On introducing the basic principle of interest vector model and the technology of user coordination filtration recommendation, a kind of efficient algorithm of goods recommendation based on interest vector model is discussed. This algorithm can realize goods automatic recommendation for electronic business by combining the interest vector model and the technology of coordination filtration recommendation, so it can increase the precision and quality of recommendation. Then, an experiment is given to check the efficiency of this recommendation algorithm.

Key words: interest vector model; coordination filtration; goods recommendation

基于关联规则的算法是一种通用化的方法，数据计算量大而繁杂；内容过滤主要存在同主题信息质量区分能力差、缺乏创新性和准确性3个问题；协同过滤则有冷启动、稀疏性和扩展实时性3个问题[1]。本文利用兴趣度向量模型，结合用户协同过滤推荐技术来实现电子商务中商品的自动推荐，从而进一步提高商品推荐质量。

1　基于兴趣度向量模型的用户协同过滤推荐技术

1.1　多维兴趣度向量模型[2]

多维兴趣度向量模型就是把用户兴趣从简单的对商品项目整体印象细化到用户对商品项目各个维度属性特征的兴趣度，对于枚举出来的商品项目多维度特征，用户对他们的兴趣是模糊的，即对某一属性分量除了具有肯定（值为1）和否定（值为0），还可能有模糊的选择（值在0和1之间）。

本文以某一电影商务网站中众多用户对电影评分所收集到的数据为例来说明用户的兴趣度向量模型，如表1所示。

电影项目特征的产地维度 V_1=（P_1，P_2，…，P_k），而兴趣维度 $\boldsymbol{F_1}$=（W_1，W_2，…，W_k）为产地维度 V_1 下用户的兴趣向量，W_1，W_2，…，W_k 分别对应用户对属性 P_1，P_2，…，P_k 具有的模糊兴趣度，衡量用户 U_i 对属性分量 p_j 的兴趣度方法为[3]：

收稿日期：2007-12-03

作者简介：肖满生（1968-），男，湖南邵东人，湖南工业大学高级讲师，硕士，主要研究方向为数据库和数据挖掘；
王　宏（1977-），男，湖南株洲人，湖南工业大学讲师，硕士，主要研究方向为数据库和数据挖掘.

$$CTI(U_i,P_j)=\frac{\sum R(P_j)/m}{\sum R(TotleU_i)/n}, \quad (1)$$

上式中，CTI是用户兴趣度（Customer Interest）的简称，$\sum R(P_j)$表示用户U_i评分过的具有属性分量p_j的项目得分之和，$\sum R(TotleU_i)$表示用户U_i评分过的所有项目的得分之和，m是U_i评分过的具有特征分量p_j的项目数目，n是U_i评分过的项目总数。

表1 V_1维度下电影项目属性分量及评分

Tab. 1 The properties of film project and their score based on V_1 dimensions

项目	属性分量						评分
	p_1	p_2	p_3	p_4	p_5	p_6	
门徒	1	0	0	1	0	1	5
天行者	0	0	1	1	1	1	4
生日快乐	0	1	0	0	1	0	3
双子神偷	0	1	1	1	0	0	5
刀马旦	1	0	0	0	1	1	1
八旗子弟	1	0	0	0	1	1	2
九龙冰室	0	1	0	1	1	0	3

利用公式（1）计算出任一评分用户U_i对某一属性分量p_j的兴趣度之后，还可进一步计算用户U_i对某一属性分量的相对兴趣度：

$$CTRI(U_i,P_j)=\frac{CTI(U_i,P_j)\times n}{\sum_{i=1}^{n}CTI(U_i,P_j)}, \quad (2)$$

式中，用户相对兴趣度$CTRI$（Customer Relative Interest）表示用户U_i相对全体用户群对属性分量p_j的兴趣程度，我们可以得到U_i对项目特征维度V_1=（P_1，P_2，…，P_k）中各个属性的相对兴趣度$CTRI(U_i, P_1)$，$CTRI(U_i, P_2)$，…，$CTRI(U_i, P_k)$。分别替换成W_1，W_2，…，W_k，便得到了用户U_i的兴趣度向量$\boldsymbol{F_1}$=（W_1，W_2，…，W_k）。同理可以得到用户的其它维度的兴趣向量$\boldsymbol{F_2}$，…，$\boldsymbol{F_k}$，再按重要程度进行加权求和，得到多维用户兴趣向量模型，即：$U=a_1F_1+a_2F_2+\cdots+a_nF_n$， (3)

其中各个维度的兴趣度向量的权值具有关系：

$a_1+a_2+\cdots+a_n=1$。

1.2 用户协同过滤推荐技术

协同过滤分析用户的兴趣所在，在用户群组中找到与指定用户兴趣相似的用户，综合这些相似用户对某一信息的评价，形成系统对该指定用户对此信息的喜好程度预测，用户协同过滤推荐的出发点有3个：1）用户是可以按兴趣来进行分类的；2）用户对不同信息评价包含了用户的兴趣信息；3）用户对一未知信息的评价将与其兴趣相似用户的评价相似。

与传统内容过滤相比，用户协同过滤的优点有：1）能够过滤难以进行机器自动基于内容分析的信息，如电影；2）能够基于一些复杂的、难以表达的概念（信息质量、品位）进行过滤；3）推荐的新颖性。

目前，用户协同过滤推荐算法可以分为两类：一类是全局数值算法，或称为基于内存的算法，在对某个特定用户作预测时需要对整个用户数据库进行比较计算；另一类是基于模型算法，利用用户数据产生用户模型，根据模型作预测。以上算法请读者参阅有关文献[4]，此处不再赘述。当然，要想获得满意的推荐效果，首先必须要得到准确的用户信息，这是协同过滤的本质所在。

1.3 多维兴趣度向量模型和用户协同推荐策略结合

多维兴趣度向量模型和用户协同推荐策略的结合即是从协同过滤推荐技术出发，为目标用户寻找与其兴趣度最相似的“邻居”，基于1个维度的邻居相似度的计算方法为[5]：

$$Sim(D,Q)\_F_1=\frac{\sum_{i=1}^{k}W_{D_i}\cdot W_{Q_i}}{\sqrt{\sum_{i=1}^{k}(W_{D_i})^2\cdot\sum_{i=1}^{k}(W_{Q_i})^2}}, \quad (4)$$

该公式根据用户之间兴趣特征向量夹角余弦值来计算用户之间的匹配度，数值越大说明他们的兴趣越相似，基于多维兴趣度相似邻居的计算方法为：

$$Sim(D,Q)=\sum_{i}^{n}a_i\cdot Sim(D,Q)\_F_i, \quad (5)$$

该方法即在公式（4）的基础上加权求和，a_1，a_2，…，a_n是和为1的一组权值，其大小可以根据经验按兴趣特征的重要性分配。

在电子商务网站上，商品项目评分矩阵是高度稀疏矩阵，利用多维兴趣度向量模型，可以根据不同用户的兴趣度的差异为其寻找相似的用户群组，进一步用协同过滤的方法作出评分预测和推荐，降低了数据稀疏的影响。另外，该向量模型还可以使大量的工作离线完成，在线处理的主要工作只是计算目标用户与用户群组中的其他用户的兴趣相似度以及预测，有效解决了推荐系统的实时性和扩展性问题。

2 基于兴趣度向量模型的用户协同过滤推荐算法描述

利用兴趣度向量模型为用户推荐商品项目的算法如下：第1步，从推荐系统的输入模块中收集用户对项目的评价信息，并建立商品项目特征文件；第2步，建立用户兴趣模型，根据评价信息以及项目特征文件，计算每一个用户对商品项目各个维度的各个属性分量的相对兴趣度，得到每个用户的多维兴趣向量模型；第3步，搜索兴趣相似的邻居，这一步骤可以在线进行，当目标用户访问网站的时候，推荐系统计算其他用户兴趣度向量与目标用户的夹角余弦值，挑选符合要求的用户作为其最近邻居；第4步，预测

和推荐，按照协同过滤的推荐原理，参照目标用户最近邻居的兴趣，根据邻居们对项目的评分，用公式 $P_{u,i}=\bar{r}_u+\sum_{v=1}^{n} Sim(u,v)\times(r_{v,i}-\bar{r}_v)$[6]，将预测得分最高的项目作为推荐系统的输出。算法描述如图 1 所示。

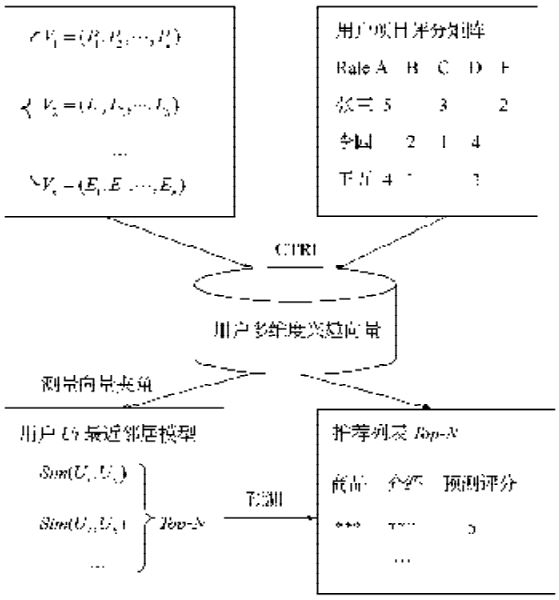

图 1　用户兴趣度向量模型的推荐算法

Fig. 1　The recommendation algorithms of the user interest vector mode

3　推荐算法的有效性验证

为了验证上述推荐算法的有效性，采用鸿波电影网站（http://tkww79.hbol.net）的用户（超过 10 000 人）对电影（超过 2 000 部）评分所涉及的数据为数据集，该数据集包含 3 个数据文件，分别为：

1）评分信息文件 rating.dat，包括字段：User_ID（评分用户标识）、Move_ID（被评价的电影标识）、Rating（评分）和 Timestamp（评分时间）；

2）电影信息文件 Movies.dat，包括字段 Movie_ID（影片标识）、Title（题目）、Genres（风格类别），其中 Genres 是非常重要的信息，可以作为电影特征的 1 个维度，也即用户兴趣的体现；

3）用户信息文件 User.dat，包括字段 User_ID（用户标识）、Gender（性别）、Age（年龄）、Occupation（职业）、Zip_code（身份证号码）。

要验证基于用户兴趣向量的协同过滤推荐算法的有效性，从数据集中挑选了 2 组实验数据，其中第 1 组包含 5 000 条评分数据，122 个用户和 1 030 部电影，数据稀疏等级：1−5 000/（122 × 1 030）=0.960 2；第 2 组包含 10 000 条评分数据、183 个用户、1 030 部电影，数据稀疏等级：1−10 000/（183 × 1 030）=0.946 9，显然，第 1 组比第 2 组数据更稀疏，实验在 2 组密度不同的数据集上比较该策略与传统协同过滤推荐方法的准确性。

在此基础上从第 1 组实验数据中随机抽取 25 个用户、在第 2 组抽取 23 个用户假定为推荐目标客户群组，他们的 80 % 评分信息作为已知，利用其余 20 % 的评分信息来测试预测推荐的精度。评价系统推荐精度采用统计度量方法中的平均绝对偏差 *MAE*（men absolute error）来对推荐质量进行度量，该方法通过计算预测的用户评分与实际的用户评分之间的偏差度量预测的准确性，*MAE* 越小，推荐质量越高。具体对比如图 2 所示。

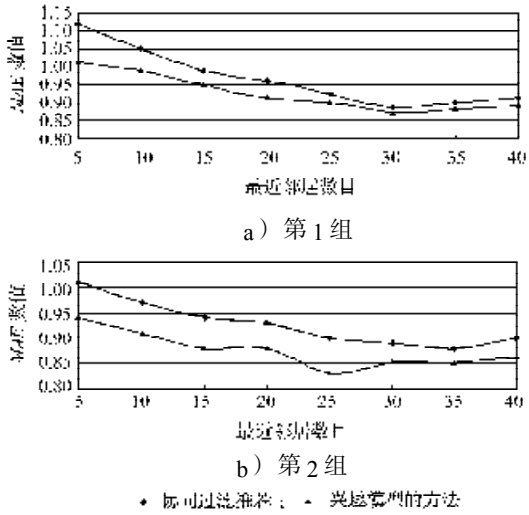

图 2　推荐方法 *MAE* 对比

Fig. 2　Comparison *MAE* of the recommendation method

从图 2 可以看出，通过对不同数据的最近邻居、稀疏程度不同的数据集分别进行多次计算，可以得出，基于兴趣度向量模型的推荐方法比传统的协同过滤推荐有更小的预测偏差，从而提高了推荐质量。

参考文献:

[1] 张云涛，龚　玲. 数据挖掘原理与技术[M]. 北京：电子工业出版社，2004.

[2] 黄萱菁，夏迎炬，吴立德. 基于向量空间模型的文本过滤系统[J]. 软件学报，2003，14(3)：438.

[3] 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.

[4] 邢春晓，高凤荣，战思南，等. 适应用户兴趣变化的协同过滤推荐算法[J]. 计算机研究与发展，2007，44(2)：296−301.

[5] 赵晓煜，丁延玲. 基于顾客交易数据的电子商务推荐方法研究[J]. 现代管理科学，2006(3)：93−94.

[6] David Hand，Heikki Mannila，Padhraic Smyth. 数据挖掘原理[M]. 北京：机械工业出版社，2003.

（责任编辑：罗立宇）