

# 协同过滤推荐算法综述

马宏伟<sup>1</sup>, 张光卫<sup>2</sup>, 李 鹏<sup>3</sup>

<sup>1</sup>(山东建筑大学 计算机科学与技术学院, 山东 济南 250101)

<sup>2</sup>(北京航空航天大学 软件开发环境国家重点实验室, 北京 100083)

<sup>3</sup>(哈尔滨工业大学 深圳研究生院 信息安全中心, 广东 深圳 518055)

E-mail: mahongwei@sdjzu.edu.cn

**摘要:** 推荐系统是电子商务系统最重要的技术之一, 协同过滤推荐是目前应用最广泛和最成功的推荐系统. 介绍协同过滤推荐算法的基本思想和最新研究进展, 分析目前出现的代表性算法. 总结协同过滤推荐算法中的关键问题和相关解决方案, 比如相似性比较, 数据稀疏性问题, 推荐的实时性, 推荐策略, 评估方法等, 同时也对比分析各种方法的优缺点. 最后介绍协同过滤推荐算法需要进一步解决的问题和可能的发展方向.

**关键词:** 推荐系统; 协同过滤推荐算法; 稀疏性; 扩展性

**中图分类号:** TP311

**文献标识码:** A

**文章编号:** 1000-1220(2009)07-1282-07

## Survey of Collaborative Filtering Algorithms

MA Hong-wei<sup>1</sup>, ZHANG Guang-wei<sup>2</sup>, LI Peng<sup>3</sup>

<sup>1</sup>(Department of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, China)

<sup>2</sup>(State key Laboratory of Software Development Environment, Beijing University of Aeronautics and Astronautics, Beijing 100083, China)

<sup>3</sup>(Information Security Center of Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China)

**Abstract:** Recommender system in E-commerce analyzes preferences of users and presents recommendations, offering personalized purchase service. This paper presents an overview of the field of collaborative filtering recommender systems and describes main techniques applied in the current generation of collaborative filtering algorithm. This paper also describes various limitations of current recommendation methods and discusses the possible ways that can improve recommendation capabilities and make recommender systems applicable to an even larger range of applications.

**Key words:** recommender system; collaborative system; sparsity; scalability

## 1 引言

电子商务的飞速发展将人类带入了网络经济时代, 面对大量的商品信息, 用户(消费者)往往难以发现最需要或最适合的商品. 电子商务系统会形成海量的交易数据, 如何从中挖掘和发现有用的知识以使得交易更加高效成为一个有意义的研究课题. 消费者希望电子商务系统具有一种类似采购助手的功能来帮助其选购商品, 它能够自动地把用户可能最感兴趣的商品推荐出来<sup>[1]</sup>. 电子商务推荐系统正是针对以上问题和需求产生的, 为了提供精确而又快速的推荐, 研究者提出了多种推荐算法<sup>[2-6]</sup>, 其中协同过滤推荐算法是应用最为成功的一种<sup>[2,7-11]</sup>.

协同过滤这一概念首次于1992年由Goldberg、Nicols、Oki及Terry提出<sup>[13]</sup>, 应用于Tapestry系统, 该系统仅适用较小用户群(比如, 某一个单位内部), 而且对用户有过多要求(比如, 要求用户显式的给出评价). 作为协同过滤推荐系统的雏形, Tapestry展示了一种新的推荐思想, 但存在许多技术上的不足. 其后, 出现了基于评分的自动协同过滤推荐系统, 例如推荐新闻和电影的GroupLens<sup>[14]</sup>. 目前, 许多电子商务网站

都已经使用了推荐系统, 如Amazon、CDNow、Drugstore和Moviefinder等<sup>[15]</sup>.

目前主要有两类协同过滤推荐算法: 基于用户的协同过滤推荐算法<sup>[7,8]</sup>和基于项目的协同过滤推荐算法<sup>[9-12]</sup>. 基于用户的协同过滤推荐算法基于这样一个假设, 即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. 算法根据目标用户的最近邻居(最相似的若干用户)对某个项目的评分逼近目标用户对该项目的评分<sup>[1,9,8]</sup>. 基于项目的协同过滤推荐算法认为, 用户对不同项目的评分存在相似性, 当需要估计用户对某个项目的评分时, 可以用用户对该项目的若干相似项目的评分进行估计<sup>[1,9-12]</sup>.

本文对协同过滤推荐算法国内外的研究现状进行了总结, 并对协同过滤推荐算法中的关键技术和存在问题进行了详细的分析, 文章最后对协同过滤推荐的发展进行了展望.

## 2 相似性比较方法

相似性计算是协同过滤推荐算法中最关键的一步, 传统的相似度计算方法有以下三种<sup>[1,7-12]</sup>:

### 2.1 余弦相似性

收稿日期: 2008-03-17 基金项目: 国家自然科学基金项目(60496323)资助. 作者简介: 马宏伟, 男, 1969年生, 博士, 副教授, 研究方向为协同计算; 张光卫, 男, 1970年生, 博士研究生, 研究方向为电子商务; 李 鹏, 男, 1977年生, 硕士研究生, 研究方向为电子商务.

把用户评分看作  $n$  维项目空间上的向量,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  $n$  维项目空间上的评分分别表示为向量  $i, j$ , 则用户  $i$  和用户  $j$  之间的相似性为:

$$sim(i, j) = \cos(i, j) = \frac{i \cdot j}{\|i\| \|j\|} \tag{1}$$

2.2 修正的余弦相似性

余弦相似性度量方法中没有考虑不同用户的评分尺度问题,修正的余弦相似性度量方法通过减去用户对项目的平均评分来改善上述缺陷,设经用户  $i$  和用户  $j$  共同评分的项目集合用  $I_{ij}$  表示,  $I_i$  和  $I_j$  分别表示经用户  $i$  和用户  $j$  评分的项目集合,则用户  $i$  和用户  $j$  之间的相似性为:

$$sim = \frac{\sum_{c \in I_{ij}} (R_{i,c} - \bar{R}_i) (R_{j,c} - \bar{R}_j)}{\sqrt{\sum_{c \in I_i} (R_{i,c} - \bar{R}_i)^2} \sqrt{\sum_{c \in I_j} (R_{j,c} - \bar{R}_j)^2}} \tag{2}$$

其中,  $R_{i,c}$  表示用户  $i$  对项目  $c$  的评分,  $\bar{R}_i$  和  $\bar{R}_j$  分别表示用户  $i$  和  $j$  对项目的平均评分.

2.3 相关相似性

设经用户  $i$  和用户  $j$  共同评分的项目集合用  $I_{ij}$  表示,则用户  $i$  和用户  $j$  之间的相似性  $sim(i, j)$  通过 Pearson 相关系数度量:

$$sim = \frac{\sum_{c \in I_{ij}} (R_{i,c} - \bar{R}_i) (R_{j,c} - \bar{R}_j)}{\sqrt{\sum_{c \in I_i} (R_{i,c} - \bar{R}_i)^2} \sqrt{\sum_{c \in I_j} (R_{j,c} - \bar{R}_j)^2}} \tag{3}$$

其中,  $R_{i,c}$  表示用户  $i$  对项目  $c$  的评分,  $\bar{R}_i$  和  $\bar{R}_j$  分别表示用户  $i$  和  $j$  对项目的平均评分.

总体来讲,三种方式均为基于向量的相似度计算方式,进行对象属性之间的严格匹配.

余弦相似性度量方法把用户评分看作一个向量,用向量的余弦夹角度量用户间的相似性,然而没有包含用户评分的统计特征;修正的余弦相似性方法在余弦相似性基础上,减去了用户对项目的平均评分,然而该方法更多体现的是用户之间的相关性而非相似性,相关性和相似性是两个不同的概念,相似性反映的是聚合特点,而相关性反映的是组合特点<sup>[16,17]</sup>;相似相关性方法,依据双方共同评分的项目进行用户相似性评价,如果用户间的所有评分项目均为共同评分项目,那么相似相关性和修正的余弦相似性是等同的.用户对共同评分项目的评分确实能很好地体现用户的相似程度,但由于用户评分数据的极端稀疏性,用户间共同评分的项目极少,使得相似相关性评价方法实际不可行.文献<sup>[17]</sup>给出一种定性知识指导的相似性评价方法.

随着电子商务系统规模的扩大,用户数目和项目数据急剧增加,使得用户-项目矩阵的极端稀疏性<sup>[10]</sup>,导致用户最近邻居和项目最近邻居的计算准确性降低,使得推荐系统的推荐质量急剧下降<sup>[7-11]</sup>.针对这种情况研究者提出多种解决方法,其中包括矩阵填充、矩阵降维等技术.

3 用户-项目矩阵稀疏性问题及解决办法

3.1 矩阵填充技术

最简单的填充办法就是将用户对未评分项目的评分设为一个固定的缺省值,或者设为其他用户对该项目的平均评分<sup>[2]</sup>.然而用户对未评分项目的评分不可能完全相同,这种办法不能从根本上解决稀疏性问题.采用预测评分的方法填充用户-项目矩阵能够产生较理想的推荐效果,主要有以下几类:

3.1.1 BP 神经网络<sup>[18]</sup>.

BP 神经网络对复杂的输入输出关系有比较强大的学习和建模能力,能够有效地处理非完整信息.图 1 表示通过某用户对项目  $I_1, I_2, I_3, I_4, I_5$  的评分来预测该用户对  $I_6$  的评分. BP 神经网络是一个 3 层网络,分别为输入层、隐含层和输出层.激活函数使用 Logistic Function:

$$F(\text{net}) = a + \frac{b}{1 + e^{-d \times \text{net}}} \tag{4}$$

其中,  $a, b, d$  为常数,  $\text{net}$  为神经元的加权输入.

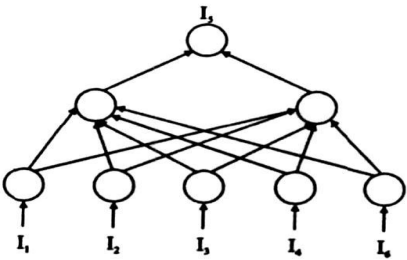


图 1 BP 神经网络结构图  
Fig. 1 Structure of neural networks

BP 神经网络把用户对各个项目的评分看作训练样本,分别输入到输入层的各个单元中;这些单元经过加权,输出到隐含层的各个单元;隐含层的加权输出再经过一次加权作为输出层的单元输入;最后由输出层产生给定样本的预测值.

这种矩阵填充技术对噪声数据有较强的承受能力<sup>[19]</sup>,可以有效降低用户-项矩阵的稀疏性,达到提高推荐精度的目的.然而, BP 算法的缺点为存在随着训练时间的增加,收敛速度有变慢的趋势<sup>[20]</sup>,以致会延长最近邻居的查找时间.文献<sup>[20]</sup>对 BP 算法进行了改进,有效缓解了上述问题.

3.1.2 Naive Bayesian 分类方法

Naive Bayesian 分类方法基于概率模型进行分类<sup>[21-24]</sup>,可以使用该方法估算一个实例属于某一类的概率,在得到某一个项目所属的分类之后,可以利用此分类中其他项目的评分情况来预测未评分项目的评分,从而可以填充用户-项目矩阵,降低稀疏性.

给定项目分类  $c_i$ , Naive Bayesian 分类方法从学习数据集中获取每个类别属性变量  $X_k$  的条件概率,然后在类别属性值  $X_1, \dots, X_n$  的情况下用 Bayesian 公式计算  $C$  的概率,如下所示:

$$P(C = c_i | X_1 = x_1, \dots, X_n = x_n) \tag{5}$$

如果认为这些变量是相互独立的,那么可以按照如下的公式计算概率:

$$P(C = c_i | X_1 = x_1, \dots, X_n = x_n) \propto$$

$$P(C=c_i) \prod_{k=1}^n P(X_k=x_k | C=c_i) \quad (6)$$

由于  $P(C=c_i)$  和  $P(X_k=x_k | C=c_i)$  都是从学习集中得到的, 因此上述方法很适合进行机器学习. 计算得到概率最高的类就是我们要找的类.

### 3.1.3 基于内容的预测<sup>[25-31]</sup>

基于内容的预测又称基于属性的预测或基于语义的预测, 该方法根据项目的属性联系以及项目所处的地位、相互关系和项目元信息等内容计算项目之间的内容相似性, 而不依赖于用户对项目的评分. 一种简化了的计算方法为:

令  $i_p = \{p_1, p_2, \dots, p_m\}$  和  $i_q = \{q_1, q_2, \dots, q_m\}$  表示两个项目的属性向量,  $i_p$  和  $i_q$  的内容相似性为:

$$S(i_p, i_q) = \frac{\sum_{i=1}^m \sum_{j=1}^n S(p_i, q_j)}{|i_p| \cdot |i_q|} \quad (7)$$

其中,  $S(p_i, q_j)$  为  $i_p$  和  $i_q$  某两个属性之间的内容相似性.

得到项目之间的内容相似性后, 选择与目标项目相似性最大的若干个项目进行评分预测, 用预测评分填充用户-项目矩阵中的空项, 降低其稀疏性. 由于不同类别的项目之间在属性描述上有较大差别, 因此基于语义的方法无法计算跨类别的项目之间的相似性, 也就无法进行跨类别的评分预测<sup>[31]</sup>. 另外基于语义的相似性计算需要提取项目的属性特征, 涉及到领域知识, 应用面较窄.

## 3.2 矩阵降维技术-奇异值分解

通过降低用户-项目矩阵的维数解决矩阵的稀疏性问题, 奇异值分解 (Singular Value Decomposition, SVD) 是一种矩阵分解技术, 它深刻揭露了矩阵的内部结构, 它可以将一个  $m \times n$  (假设  $m \geq n$ ) 的矩阵  $R$  分解为三个矩阵  $U, S, V$ <sup>[22]</sup>, 大小分别为  $m \times m, m \times n, n \times n$ .

得到用户-项目矩阵的SVD分解后, 分别将  $U, S, V$  降为  $k$  维的方阵, 得到  $U_k, S_k$  和  $V_k$ <sup>[32]</sup>. 用户  $u$  对项目  $i$  的评分预测公式为:

$$P_{u,i} = \overline{R_u} + U_k \times \overline{S_k(a)} \times \overline{S_k} V_k(i) \quad (8)$$

其中,  $\overline{R_u}$  是用户在他已打分项目上的平均评分值,  $k$  值的大小会影响评分预测的效果, 如果  $k$  值太小, 则会丢失原始评分矩阵中的重要结构; 如果  $k$  值太大, 则失去了降维的意义. 针对不同的原始评分矩阵, 通过实验来选取  $k$  值, 以达到最佳的降维效果.

协同过滤推荐系统中SVD的优势主要体现在<sup>[33]</sup>: 用户-项目矩阵的稀疏性问题得到很好的解决; 对用户-项目矩阵降维后, 运算复杂度大大降低, 系统的扩展性得到提升; 用户间和项目间的潜在关系将得到更好的发掘, 有利于提高推荐精度. SVD方法的缺点为: 降维会导致用户-项目矩阵中的信息丢失<sup>[33,34]</sup>, 有的情况下会影响推荐精度, 通过选取合适的保留维数  $k$ , 可以在一定程度上减小这种影响.

总的来说, SVD方法不仅能够解决矩阵稀疏性问题, 而且对于系统的扩展性和推荐精度的提高也有作用.

另外一种解决稀疏性问题的方法是在系统中嵌入一种

半智能化 (semi-intelligent filtering) 的智能体<sup>[36,37]</sup> (agent) 或类似的自动关联评分系统, 使用句法特征来预测用户对未评分项目的评分, 这种方法能够在一定程度上解决稀疏性问题, 但同时会影响系统的推荐速度, 应用领域会受到限制.

## 4 冷启动问题

考虑如下两种情况:

1) 在 User-based 系统中, 对于一个新的用户来说, 系统中没有该用户的任何购买信息记录, 因此无法找到其最近邻居, 从而无法进行推荐.

2) 在 Item-based 系统中, 当系统中加入一个新的项目时, 该项目没有评分记录, 无法找出其最近邻居并进行推荐或评分预测.

协同过滤推荐系统中存在的这种问题被称为冷启动问题<sup>[30,37,38]</sup>. 为了解决冷启动问题, 普遍采用基于内容的最近邻居查找技术<sup>[1,26,28,31]</sup>, 其基本思想是:

1) 利用聚类技术将用户按照属性相似性聚类, 从项目属性的角度找到新项目的最近邻居;

2) 用新项目  $k$  的所有最近邻居的平均评分来代替已有评分的平均值  $\overline{R_k}$ .

## 5 推荐速度

电子商务系统中, 由于项目在一定时期内通常是相对稳定的<sup>[2,28-30]</sup>, 项目相似性的计算可以离线进行, 这就使得与基于用户的协同过滤推荐系统相比, 基于项目的协同过滤推荐系统的运行效率较高.

随着用户数和项目数的增多, 协同过滤推荐算法的计算量也不断增大<sup>[2,28-30]</sup>. 通常采用聚类技术提高推荐速度, 因为使用聚类技术可以大大缩小用户或项目的最近邻居搜索范围, 从而提高推荐的实时性<sup>[39-44]</sup>.

### 5.1 EM (Expectation-Maximization) 算法

EM算法通过估计用户或项目属于某一类的概率对用户或项目进行聚类. 假设  $P_k$  为任一用户属于用户类别  $k$  的概率,  $P_i$  为任一项目属于项目类别  $i$  的概率,  $P_{ki}$  为  $k$  中的用户与  $i$  中的项目有关联的概率. EM算法首先估计这些参数, 然后进行聚类. 在实际的聚类中, 不同的用户可能会喜欢同一个项目, 根据EM算法, 这样的项目可能会同时出现在两个不同的聚类中. 如果要求每个用户或项目只能属于一个用户分类或项目分类, EM算法就不再适用.

### 5.2 k-means 聚类算法

以项目聚类为例, k-means聚类算法通过用户对项目评分的相似性对项目进行聚类并生成相应的聚类中心, 然后计算目标项目与各聚类中心的相似度, 选出与目标项目相似度最高的  $k$  个聚类中心对应的聚类, 在这  $k$  个聚类中搜索目标项目的最近邻居, 从而达到在尽量少的项目空间中找到目标项目的大部分最近邻居<sup>[31-33]</sup>. K-means聚类算法的优点在于不同聚类中的项目之间有明显的区别, 而且算法的扩展性

相对较好. 缺点是聚类数目  $k$  需要事先给定而且不同的应用中  $k$  值是不同的, 难于选取; 另外初始聚类中心是随机选取的, 对于同一组数据, 可能因为初始聚类中心的不同而产生不同的聚类结果<sup>[42]</sup>.

5.3 Gibbs Sampling 方法<sup>[42]</sup>

Gibbs Sampling 方法与 EM 算法类似, 首先估计三个概率参数:  $P_k$ 、 $P_i$  和  $P_{ki}$ , 然后进行聚类. 不同的是 Gibbs Sampling 方法基于 Bayesian 模型. Gibbs Sampling 算法有较好的聚类效果和很强的扩展性, 但是其算法复杂度较大. 聚类过程相对比较耗时, 但是可以离线进行, 而且目标项目的最近邻居搜索范围缩小到几个聚类中, 远远小于整个项目空间.

5.4 模糊聚类<sup>[42-44]</sup>

模糊聚类与聚类的区别在于前者不需要预先给定聚类的数目, 而是通过一定的阈值来确定对象的相似类别. 模糊聚类利用模糊等价关系将给定的对象分为一些等价类, 并由此得到与关系对应的模糊相似矩阵, 该模糊相似矩阵满足传递性. 根据相似矩阵求出其传递关系的闭包, 然后在传递关系的闭包上实现分类, 步骤如下:

(1) 计算两个项目之间的相似度, 建立模糊相似矩阵  $M_{m \times m}$ , 其中当  $i=j$  时, 令  $M_{ij}=1$ ; 当  $i \neq j$  时, 令  $M_{ij}$  为项目  $i$  和  $j$  之间的属性相似度.

(2) 计算相似矩阵  $M_{m \times m}$  的传递包, 得到等价矩阵.

(3) 设置不同的阈值  $\lambda$  确定相应的截集, 即对  $M$  矩阵的模糊相似关系图进行模糊聚类: 计算  $[X]=\{y \mid M(x, y) \leq \lambda\}$ , 集合  $[X]$  即为模糊聚类的等价类, 也就是对应相似关系图上的连通分量, 每个连通分量就是对应此阈值的某个模糊等价类.

模糊聚类过程可以离线进行, 不会给推荐系统的实时性带来负担. 同时模糊聚类对于解决数据稀疏性带来的冷启动问题也有很好的效果<sup>[41, 44]</sup>.

6 推荐策略

6.1 平均加权策略

目前大多数协同过滤推荐系统都采用平均加权策略产生推荐<sup>[1, 10-13, 25-28]</sup>, 目标用户  $u$  对未评分项目  $i$  的预测评分为:

$$P_{u,i} = \bar{R}_u + \frac{\sum sim(u,n) \times (R_{n,i} - \bar{R}_n)}{\sum sim(u,n)} \quad (9)$$

其中,  $sim(u, n)$  为用户  $u$  和用户  $n$  的相似度,  $R_{n,i}$  为最近邻居集中的用户  $n$  对项目  $i$  的评分,  $\bar{R}_u$ 、 $\bar{R}_n$  分别为用户  $u$  和用户  $n$  对项目的平均评分. 平均加权策略在产生推荐的时候综合考虑了用户对所有项目的评分情况. 在用户评价过的项目数较多时, 这种方法是合理的而且实验证明有较好的推荐效果, 当用户评价过的项目数较少时, 个别项目的评分就会对平均评分产生较大影响, 这种情况下平均评分无法反映用户对大多数项目的评分情况.

6.2 基于评分频度的推荐策略

电子商务中用户评分通常为离散值, 比如  $\{1, 2, 3, 4, 5\}$ , 基于评分频度的推荐策略首先用统计的方法计算最近邻居集

中用户给出的各种评分的出现频率, 然后将评分频率最高的那个分值作为目标用户的预测评分. 该推荐策略用户对大多数邻居的评分情况, 实验证明, 在最近邻居数较少时, 推荐效果优于平均加权策略<sup>[17]</sup>.

7 评估方法

数据源: 目前大多数关于协同过滤推荐算法的实验数据都取自 MovieLens 站点 (<http://movielens.umn.edu>)<sup>[9, 10, 33, 37]</sup>, 该站点是一个基于 Web 的研究型推荐系统, 注册用户必须至少对它所拥有的影片中的 15 部进行评价才可以使用该系统. 另外, 还有 EachMovie 数据集<sup>[32]</sup> (由 DEC 提供, 目前已不提供下载)、MRS (Movie Recommendation System)<sup>[32]</sup>等.

推荐质量的评价标准主要有统计精度度量和决策支持精度度量方法两类. 统计精度度量方法中常用的是平均绝对偏差 MAE<sup>[12, 19, 25, 34]</sup> (Mean Absolute Error); 决策支持精度度量方法中主要有召回率 (Recall)<sup>[25]</sup>、准确率 (Precision)<sup>[37]</sup> 及 ROC (Receiver Operating Characteristic)<sup>[49]</sup> 等三种方法.

1) 平均绝对偏差 (MAE) 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<sup>[37-41]</sup>. 假设预测的用户评分集合为  $\{p_1, p_2, \dots, p_N\}$ , 对应的实际评分集合为  $\{q_1, q_2, \dots, q_N\}$ , 则 MAE 可由下式计算:

$$MAE = \frac{\sum_{i=1}^N |p_i - q_i|}{N} \quad (10)$$

2) 召回率 (Recall) 反映了待推荐项目被推荐的比率:

$$recall = \frac{|\text{test} \cap \text{top-N}|}{|\text{test}|} \quad (11)$$

其中 test 表示测试数据集中的项目数量, top-N 表示系统推荐给用户的  $N$  个项目.

3) 推荐准确率 (Precision) 表示算法推荐成功的比率.

$$precision = \frac{|\text{test} \cap \text{top-N}|}{N} \quad (12)$$

其中 test 表示测试数据集中的项目数量, top-N 表示系统推荐给用户的  $N$  个项目. 2) 中所述的召回率 Recall 和推荐准确度 Precision 是一对矛盾, 当 top-N 个数增加时, 召回率 (Recall) 升高而推荐准确率 (Precision) 下降. 一个推荐结果的召回率和推荐精度同样重要, 故而综合考虑算法这两方面的性能是比较科学的. 如下所示的参数 F1 采用相同的权重将召回率和推荐精度结合起来, 在二者之间找到最佳平衡点<sup>[47-49]</sup>.

$$F1 = \frac{2 * recall * precision}{recall + precision} \quad (13)$$

4) ROC (Receiver Operating Characteristic) 有两个指标<sup>[48]</sup>: Sensitivity 和 Specificity. 前者为任选一个用户喜欢的项目, 该项目被系统推荐的概率, 后者为任选一个用户不喜欢

的项目,该项目未被系统推荐的概率.在协同过滤推荐系统中,系统对目标项目的预测分值和用户对该项目的实际评分

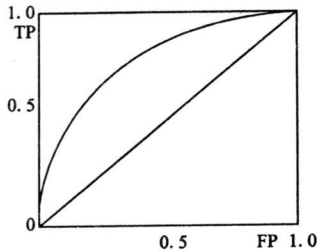


图2 ROC 曲线

Fig.2 ROC curve

之间存在一个差值.设定一个域值,差值小于域值的,认为是用户喜欢的项目,差值大于域值的,认为是用户不喜欢的项目.图2所示为ROC曲线,纵坐标为Sensitivity,横坐标为1-specificity.

Sensitivity 值在0到1之间变化,给定一个域值,就对应图中左上角的曲线上的一个点.图中穿过原点和(1,1)点的直线表示一个随机推荐的系统,即任选一个用户喜欢的项目,该项目被系统推荐的概率为0.5.由图2可知,曲线越向左上角靠近,说明推荐系统的推荐精度越高,反之则越低.

## 8 结论及展望

本文介绍了协同过滤推荐算法的主要思想及算法面临的主要问题,总结了近期本领域内的研究进展,对比分析了针对数据稀疏性和实时性问题提出的多种解决方法.最后对推荐算法的推荐策略及算法评价方法进行了较详细的介绍.

研究者提出了多种方法来解决协同过滤推荐算法面临的数据稀疏性和推荐实时性问题,一个明显的趋势是其他领域的一些技术正被应用到推荐系统中来,例如利用BP神经网络技术来解决数据稀疏性问题.这些方法在一定程度和某些场合下有效缓解了数据稀疏性问题并提高了推荐速度,然而随着电子商务的进一步发展,这些问题将进一步凸现.将领域外的一些技术与推荐算法结合起来解决面临的问题将是一个有意义的研究方向.另外,随着各种各样的推荐算法层出不穷,对算法的评价标准的研究也显得越来越重要.

## References:

- [1] Arwar B, Karypis G, Konstan J, et al. Analysis of recommendation algorithms for E-commerce[C]. In: Processing of 2nd ACM Conference on Electronic Commerce, 2000, 158-167.
- [2] You Wen, Ye Shu-sheng. A survey of collaborative filtering algorithm applied in E-commerce recommender system[J]. Computer Technology and Development, 2006, 16(9): 70-72.
- [3] Wang Zhir-mei, Yang Fan. P2P recommendation algorithm based on hebbian consistency learning[J]. Computer Engineering and Applications, 2006, 42(36): 110-113.
- [4] Wang Wei-ping, Liu Ying. Recommendation algorithm based on customer behavior locus[J]. Computer Systems & Applications, 2006, 15(9): 35-38.
- [5] Gao Jing, Ying Ji-kang. A recommendation system based on artificial immune system[J]. Computer Technology and Development, 2007, 17(5): 180-183.
- [6] Xu Yr-feng, Xu Qing-yun, Liu Xiao-ping. A recommender arithmetic based on the order of time[J]. Computer Systems & Applications, 2006, 15(10): 22-25, 29.
- [7] Y Chuan, X Ji-e-ping. Recommendation algorithm combining the user-based classified regression and the item-based filtering[C]. In: Processing of the International Conference on Electronic Commerce, Proceedings - the new E-commerce: Innovations for Conquering Current Barriers, Obstacles and Limitations to Conducting Successful Business on the Internet, 2006, 574-578.
- [8] Breese J, Hecherman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering[C]. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI '98), 1998, 43-52.
- [9] Arwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]. In: Proceedings of the 10th International World Wide Web Conference, 2001, 285-295.
- [10] Kim B M, Li Q, Park C S, et al. A new approach for combining content-based and collaborative filters[J]. Journal of Intelligent Information System, 2006, 27(1): 79-91.
- [11] Karypis G. Evaluation of item-based top-n recommendation algorithms[C]. In: Proc. of the Tenth International Conference on Information and Knowledge Management, 2001, 247-254.
- [12] Deng Ai-lin, Zhu Yang-yong, Shi Bai-le. A collaborative filtering recommendation algorithm based on item rating prediction[J]. Journal of Software, 2003, 14(9): 1621-1628.
- [13] Goldberg D, Nichols D, Oki B M, et al. Using collaborative filtering to weave an information tapestry[J]. Communications of the ACM, December, 1992, 35(12): 61-70.
- [14] Resnick P, Iacovou N, Suchak M, et al. GroupLens: an open architecture for collaborative filtering of netnews[C]. In: Proceedings of CSCW '94, 1994, 175-186.
- [15] Wu Yan, Shen Jie, Gu Tian-zhu, et al. Algorithm for sparse problem in collaborative filtering[J]. Application Research of Computers, 2007, 24(6): 94-97.
- [16] Liu Qun. Researches into some key aspects of Chinese-English machine translation[D]. Peking University, 2004, 47-50.
- [17] Zhang Guang-wei, Kang Jian-chu, Li He-song, et al. Context based collaborative filtering recommendation algorithm[J]. Journal of System Simulation, 2006, 8(2): 595-601.
- [18] Zhang F, Chang H Y. A collaborative filtering algorithm embedded bp network to ameliorate sparsity issue[C]. In: Proceedings of 2005 International Conference on Machine Learning and Cybernetics, 2005, 1839-1844.
- [19] Chen Gang, Liu Fa-sheng. Method for data mining based on BP neural network[J]. Computer and Modernization, 2006, 10: 20-22.
- [20] Jia Li-hui, Zhang Xi-ru. Analysis and improvements of BP algorithm[J]. Computer Technology and Development, 2006, 10(16): 101-103.

- [21] Jung K Y, Hwang H J, Kang U G. Constructing full matrix through naive Bayesian for collaborative filtering [C]. In: Computational Intelligence, pt 2, Proceedings, 2006, 4114: 1210-1215.
- [22] Jung K Y. User preference through Bayesian categorization for recommendation [C]. In: Pricai 2006: Trends in Artificial Intelligence, Proceedings, 2006, 4099: 112-119.
- [23] Robles V, Larranaga P, Menasalvas E, et al. Improvement of naive Bayes collaborative filtering using interval estimation[C]. In: Proceedings of IEEE/WIC International Conference on web Intelligence, 2003, 168-174.
- [24] Ko S J. Prediction of consumer preference through Bayesian classification and generating profile [C]. In: Proceedings of Conceptual Modeling for Novel Application Domains, 2003, 2814: 29-39.
- [25] Chedrawy Z, Abidi SSR. An adaptive personalized recommendation strategy featuring context sensitive content adaptation[C]. In: Proceedings of Adaptive Hypermedia and Adaptive Web-based Systems, 2006, 4018: 61-70.
- [26] Lu Q Q, Huang G Q. A model of collaborative filtering in E-business recommendation system based on interest measure [C]. In: Processing of the 5th Wuhan International Conference on e-business, vols 1-3 - Integration and Innovation Through Measurement and Management, 2006, 562-567.
- [27] Luo Y M, Nie G H. Research of collaborative filtering algorithm based on the semantic similarity[C]. In: Processing of the 5th Wuhan International Conference on e-business, vols 1-3 - Integration and Innovation Through Measurement and Management, 2006, 2132-2138.
- [28] Tso K, Schmidt-Thieme L. Attribute-aware collaborative filtering [J]. From Data and Information Analysis to Knowledge Engineering, 2006, 614-621.
- [29] Chedrawy Z, Abidi SSR. An item-based collaborative filtering framework featuring case based reasoning[C]. In: Proceedings of the 2005 International Conference on Artificial Intelligence, 2005, 286-292.
- [30] Ghani R, Fano A. Building recommender systems using a knowledgebase of product semantics [EB/OL]. <http://www.accenture.com/xdoc/en/services/technology/publications/recommender-ws02.pdf>, 2002-10-28/2004-02-16.
- [31] Kim B M, Li Q, Park C S, et al. A new approach for combining content-based and collaborative filters[J]. Journal of Intelligent Information Systems, 2006, 27(1): 79-91.
- [32] Zhang S, Wang W H, Ford J, et al. Using singular value decomposition approximation for collaborative filtering[C]. In: Proceedings of 7th IEEE International Conference on E-commerce Technology, 2005, 257-264.
- [33] Vozalis M G, Margaritis K G. Applying SVD on item-based filtering [C]. In: Proceedings of 5th International Conference on Intelligent Systems Design and Applications (ISDA '05) 2005, 464-469.
- [34] Charu C Aggarwal. On the effects of dimensionality reduction on high dimensional similarity search [C]. In: Proc. 12th ACM SIGMOD SIGACT SIGART Symposium on Principles of Database, 2001, 256-266.
- [35] Sarwar B M, Karypis G, Konstan J A, et al. Application of dimensionality reduction in recommender system-A case study [C]. In: ACM WebKDD 2000 Workshop, 2000.
- [36] Good N, Schafer J B, Konstan J A. Combining collaborative filtering with personal agents for better recommendations [C]. In: Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99), 1999, 439-446.
- [37] Goldberg K, Roeder T, Gupta D. Eigentaste: a constant time collaborative filtering algorithm [J]. Information Retrieval, 2001, 4(1): 133-151.
- [38] Al Mamunur Rashid, Istvan Albert, Dan Cosley, et al. Getting to know you: learning new user preferences in recommender systems [C]. In: Proceedings of the 7th International Conference on Intelligent User Interfaces, 2002.
- [39] Lee J S, Jun C H, Lee J, et al. Classification-based collaborative filtering using market basket data [J]. Expert Systems with Applications, 2005, 29(3): 700-704.
- [40] Papagelis M, Plexousakis D. Qualitative analysis of user-based and item-based prediction algorithms for recommendation agents [J]. Engineering Applications of Artificial Intelligence, 2005, 18(7): 781-789.
- [41] Min S H, Han I. Dynamic fuzzy clustering for recommender systems [C]. In: Proceedings of Advances in Knowledge Discovery and Data Mining, 2005, 3518: 480-485.
- [42] Lyle H, Ungar, Dean P, Foster. Clustering methods for collaborative filtering [C]. In: Workshop on Recommendation Systems at the Fifteenth National Conference on Artificial Intelligence, 1998.
- [43] Zhang Yu-ying, Meng Hai-dong. Research of clustering algorithm improvement in data mining techniques [J]. Journal of Inner Mongolia University of Science and Technology, 2005, 24(4): 338-341.
- [44] Zhang Hai-yan, Ding Feng, Jiang Li-hong. A collaborative filtering recommendation method based on fuzzy clustering [J]. Computer Simulation, 2005, 22(8): 144-147.
- [45] Kim T H, Yang S B. An improved neighbor selection algorithm in collaborative filtering [J]. IEICE Transactions on Information and Systems, 2005 E88D(5): 1072-1076.
- [46] Cheung K W, Tian L F. Learning user similarity and rating style for collaborative recommendation [J]. Information Retrieval, 2004, 7(3-4): 395-410.
- [47] Weng L T, Xu Y, Li Y F, et al. An improvement to collaborative filtering for recommender systems [C]. In: Processing International Conference on Computational Intelligence for Modelling, Control & Automation Jointly with International Conference on Intelligent Agents, Web Technologies & Internet Commerce, 2006, 792-795.
- [48] Yang Y, Liu X. A re-examination of text categorization methods [C]. In: Proceedings of ACM SIGIR '99 Conference, 1999, 42-49.
- [49] DeLeo J. Receiver operating characteristic laboratory

ROCLAB: software for developing decision strategies that account for uncertainty [C]. In: Proceedings of the 2nd International Symposium on Uncertainty Modeling and Analysis, IEEE Computer Society Press, College Park, MD, 1993, 318-325.

#### 附中文参考文献:

- [2] 游文, 叶水生. 电子商务推荐系统中的协同过滤推荐[J]. 计算机技术与发展, 2006, 16(9): 70-72.
- [3] 王志梅, 杨帆. 基于 Hebbian 一致性学习的 P2P 推荐算法[J]. 计算机研究与应用, 2006, 42(36): 110-113.
- [4] 王卫平, 刘颖. 基于客户行为序列的推荐算法[J]. 计算机系统应用, 2006, 15(9): 35-38.
- [5] 高静, 应吉康. 基于人工免疫系统的推荐系统[J]. 计算机技术与发展, 2007, 17(5): 180-183.
- [6] 徐义峰, 徐青云, 刘晓平. 一种基于时间序列性的推荐算法[J]. 计算机系统应用, 2006, 15(10): 22-25, 29.
- [12] 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.
- [15] 吴颜, 沈洁, 顾天竺, 等. 协同过滤推荐系统中数据稀疏问题的解决[J]. 计算机应用研究, 2007, 24(6): 94-97.
- [16] 刘群. 汉英机器翻译若干关键技术研究[D]. 北京大学, 2004, 45-50.
- [17] 张光卫, 康建初, 李鹤松, 等. 面向场景的协同过滤推荐算法[J]. 系统仿真学报, 2006, 8(2): 595-601.
- [19] 陈刚, 刘发升. 基于 BP 神经网络的数据挖掘方法[J]. 计算机与现代化, 2006, 10: 20-22.
- [20] 贾丽会, 张修如. BP 算法分析与改进[J]. 计算机技术与发展, 2006, 10(16): 101-103, 107.
- [43] 张玉英, 孟海东. 数据挖掘技术中聚类算法的改进研究[J]. 包头钢铁学院学报, 2005, 24(4): 338-341.
- [44] 张海燕, 丁峰, 姜丽红. 基于模糊聚类的协同过滤推荐方法[J]. 计算机仿真, 2005, 22(8): 144-147.