文章编号:2095-0411(2017)03-0051-09

# 个性化推荐系统综述

刘 辉,郭梦梦,潘伟强
(常州大学 商学院,江苏 常州 213164)

**摘要**:个性化推荐系统作为处理“信息超载”问题的有效工具,已经得到了广泛的研究与关注。文中对电子商务环境下的个性化推荐算法进行了归类与综述,总结了现有的各类推荐算法的优缺点与个性化推荐系统性能评价指标;电子商务个性化推荐算法具有良好的发展前景,但仍需有效解决个性化推荐系统中存在的冷启动、数据稀疏与可扩展性等问题。

**关键词**:信息超载;个性化推荐;性能评价
**中图分类号**:TP 18 **文献标志码**:A **doi**:10.3969/j.issn.2095-0411.2017.03.008

# Overview of Personalized Recommendation Systems

LIU Hui, GUO Mengmeng, PAN Weiqiang
(School of Business, Changzhou University, Changzhou 213164, China)

**Abstract**: Personal recommendation as the effective tool to tackle the “information overload” problem has attracted wide attention of many researchers. In this paper, we make some categorizations and reviews for the personalized recommendation algorithms that under the Electronic Commerce circumstances, we also summarize the strength and weakness of these algorithms as well as measures used in performance evaluation of personal recommendation systems; personalized recommendation algorithms has a perfect foreground. However, in personal recommendation systems, we still need an effective solution of cold start, data sparsity and scalability issues.

**Key words**: information overload; personal recommendation; performance evaluation

随着互联网的发展与普及,人们在享受网络资源带来极大便利的同时,也受到信息碎片化与信息超载的困扰,即人们发觉很难在海量的信息中找到满足自己真正需求的内容。虽然通过基于关键字的搜索引擎如百度、谷歌等可以满足大众的简单需求,但无法满足个性化与定制化的用户需求。因此,个性化推荐系统应运而生,成为当前解决“信息超载”问题的主流方法。例如,用户浏览商品与购买行为(如电子商务平台关于某个产品与用户的浏览记录、购买记录、购物车等)等数据实际上蕴含着用户的潜在需求与消费习惯,个性化推荐系统正是通过挖掘这些数据,捕捉用户兴趣爱好,从而将商品精准地推荐给用户。个性化推荐系统不仅满足用户的个性化需求,提升用户忠诚度,并且将潜在的用户转变为真实客户,提高商业利益。网络技术的迅猛发展拓宽了推荐系统的应用领域,推荐对象已经从最初的邮件过滤扩展到了电子商务、音乐视频网站、在线广告、社交网络、个性化阅读等领域,其中在电子商务领域的应用最为成熟。

Adomavicius 等[1]给出了推荐系统的形式化定义:将系统中所有用户集合表示为 $U$,系统中所有可推荐

**收稿日期**:2016-09-30。
**作者简介**:刘辉(1980—),男,湖南新邵人,博士,副教授,主要从事数据挖掘及生物医药数据分析研究。

的产品集合表示为 $G$($U$ 和 $G$ 的规模通常都很大,例如淘宝网拥有上千万的客户与商品等)。假设使用效用函数 $r$ 计算对象 $G$ 对 $U$ 的推荐度(例如根据卖家的信誉度与买家对产品的评价等信息),即 $r:G\times U\rightarrow R$,$R$ 是一定范围内的非负实数。对于任一用户 $u\in U$,推荐系统要找到推荐度 $R$ 最大的产品 $G^*$,如式(1):

$$G^*=\arg\max_{G\in g} r(u,g) \tag{1}$$

Resnick 等[2]于 1997 年给出了推荐系统的定义。一个完整的推荐系统由用户模型、产品模型与推荐算法 3 部分组成。用户模型用于获取、表示、存储用户的浏览行为与购买历史数据,这些数据可以通过显示与隐式 2 种方式获取。显式获取是通过用户行为(如对产品的评分、喜欢/不喜欢某个产品等)表达对产品的偏好程度,直接得到数据;隐式获取是通过系统对用户行为(如网页浏览,购买日志等)的自动追踪来获取用户对产品的兴趣偏好,间接得到数据。产品模型用于表示、存储产品的特征属性。产品不同其特征属性也不相同,在推荐文档类产品(如新闻、报纸等)时可以借助分类方法与基于内容的方法提取产品的特征属性;在推荐多媒体类产品(如视频、音乐等)时,可结合相关领域的技术与知识来抽取产品的特征属性。推荐算法作为推荐系统的关键环节,主要通过挖掘用户历史数据中蕴含的规律来获取用户的兴趣偏好与消费习惯。因此,个性化推荐系统应侧重考虑如何设计推荐算法来提高推荐的精准度[3]。尽管多种推荐算法已经被提出,但仍然不能满足用户的个性化需求,许多数据挖掘与智能信息处理领域的学者仍在不断探索。

## 1 个性化推荐算法的研究

目前主流的个性化推荐算法包括协同过滤推荐、基于内容的推荐、基于二部图的推荐、基于关联规则的推荐以及基于社交网络的推荐[4]。下面分述各类推荐算法的核心思想及各类算法的优缺点。

### 1.1 协同过滤推荐算法

Goldberg 等[5]于 1992 年提出了协同过滤(collaborative filtering)的概念,最初应用在 Tapestry System 上过滤对用户有用的电子邮件。经过近 20 年的发展,协同过滤已成为智能推荐领域的重要算法。具体地,协同过滤推荐算法是指利用大量用户与产品关联的历史数据,计算用户/产品之间的相似度,查找与目标用户相似性较高的近邻集,并通过近邻集用户对其他产品的评分来预测目标用户对该产品的潜在评分,产生推荐的产品集合[6]。协同过滤推荐算法可分为基于用户的过滤算法、基于产品的过滤算法与基于模型的推荐算法[7]。基于用户的过滤算法是指根据目标用户的偏好,找到与目标用户兴趣相似的用户群体并将该群体感兴趣的内容推荐给目标用户,为目标用户提供定制化服务;基于产品的推荐算法是指根据现有的用户行为数据,计算目标产品与用户喜欢的或已购买的产品的相似度,将相似度较高的产品推荐给用户;基于模型的方法是指根据各种机器学习的方法(如线性回归模型、朴素贝叶斯分类模型、极大熵模型等)离线建立模型,然后根据用户-产品评分矩阵,得到用户对产品的预测评分。

1)基于用户的推荐算法

该类算法根据用户对产品的评分,计算用户间的相似性并以构建的相似性矩阵为依据,估计预测评分,为用户推荐兴趣度较高的产品[8]。用户评分数据可以表示为一个 $n\times m$ 阶矩阵,$n$ 行表示共有 $n$ 个用户,$m$ 列表示共有 $m$ 个产品。$P_{i,j}$ 表示第 $i$ 个用户对第 $j$ 个产品的评分。用户评分数据矩阵见表 1。

**表 1 用户评分数据矩阵**

| | $\boldsymbol{G}_1$ | … | $\boldsymbol{G}_k$ | … | $\boldsymbol{G}_m$ |
|---|---|---|---|---|---|
| $\boldsymbol{U}_1$ | $P_{1,1}$… | … | $P_{1,k}$ | … | $P_{1,m}$ |
| … | … | … | … | … | … |
| $\boldsymbol{U}_j$ | $P_{j,1}$ | … | $P_{j,k}$ | … | $P_{1,m}$ |
| … | … | … | … | … | … |
| $\boldsymbol{U}_n$ | $P_{n,1}$ | … | $P_{n,k}$ | … | $P_{n,m}$ |

基于用户的推荐算法用于估计目标用户 $U_i$($i=1,2,\cdots,n$)对给定产品 $G_j$($j=1,2,\cdots,m$)的评分 $P_{i,j}$。该方法首先计算用户间的相似性,选取其他用户中对第 $j$ 个产品评过分的用户构成 $U_i^*$ 集,然后根据所有的 $U_k\in U_i^*$ 对第 $j$ 个产品的评分来估计用户 $U_i$ 对产品 $G_j$ 的评分[9]。该算法适用于用户相对稳定的领域,如新闻、电影与文章的推荐。尽管基于用户的推荐算法已经在智能推荐领域得到广泛应用,但该算法存在一些不足之处。以电子商务网站为例,一方面,网站产品的数量比较稳定而用户数目更新频率较高,在用户数量远大于产品数量时,计算用户间的相似性越来越耗时并占用更多内存。另一方面,基于用户的算法产生的推荐结果

可解释性较差。

2）基于产品的推荐算法

亚马逊公司于 2003 年提出了基于产品的协同过滤推荐算法[10]。该类算法不是计算用户间的相似度，而是计算目标产品与用户已购买过的或者已评过分的产品间的相似性，根据计算得到的产品-产品相似性矩阵进行评分预测，从而将用户可能感兴趣的产品加入到推荐列表中。由于电子商务平台上产品的状态相对比较稳定，因此可以通过离线的方式提前计算产品间的相似性，这样，在运行时只需要考虑用户已评分产品与其他产品的相似性，计算量大大减小。对于产品相对稳定的领域（如电子商务领域）该算法比较适用。Sarwar[11]和 Karypis G 等[12]已经证明基于产品的协同过滤推荐算法比基于用户的协同过滤推荐算法在性能上有所提升，在某些情况下（如用户数目较多时）推荐结果能更好地满足用户的个性化需求。

基于用户的推荐算法和基于产品的推荐算法涉及到用户/产品之间的相似度的计算，常用余弦相似度或修正余弦相似度、相关系数[13]等来度量用户/产品间的相似度。除此之外，许多改进的相似度计算方法已经被广泛提出并应用，如黄创光等[14]在相关研究的基础上提出了一种改良的相似性计算方法：如果用户 $U_a$ 与用户 $U_b$ 均对产品 i 进行了评分，则将产品 $i$ 加入到集合 $G'$ 中，根据设定 $\gamma$ 阀值来比较用户 $U_a$ 和 $U_b$ 共同评分的产品数目 $|G'|$，用比较结果来确定用户 $U_a$ 与用户 $U_b$ 间的相似度的大小。

$$s'(U_a,U_b)=\frac{\min(|G|,\gamma)}{\gamma}\times s(U_a,U_b) \tag{2}$$

式中：$s'$ 表示改良后的相似度；$s$ 表示用户间的相似度。从式(2)可以看到满足 $\frac{\min(|G|,\gamma)}{\gamma}\leqslant 1$，改良后的相似度 $s'(U_a,U_b)$ 的值域仍在[0,1]区间上。如果用户 $U_a$ 和 $U_b$ 共同评过分的产品较多，满足 $|I'|\geqslant\gamma$，那么 $s'(U_a,U_b)=s(U_a,U_b)$；如果共同评过分的产品较少，那么相似度量值也会相应减少。

通过余弦相似性、修正的余弦相似性和相关系数计算用户间的相似度，产生最近邻集，并通过最近邻集进行推荐，常用的推荐方法包括平均评分法、加权平均评分法，以及偏移的加权平均评分法。具体的定义如下：设 $U=(u_1,u_2,\cdots,u_n)$ 为用户的集合，$G=(g_1,g_2,\cdots,g_m)$ 为产品的集合，$r(u,g)$ 表示用户 $u$ 对产品 $g$ 的评分。

$$r(u,g)=\frac{1}{n}\sum_{k\in\bar{U}}r_{k,i} \tag{3}$$

$$r(u,g)=\frac{\sum_{k\in\bar{U}}s(u,k)r_{k,i}}{\sum_{k\in\bar{U}}|s(u,k)|} \tag{4}$$

$$r(u,g)=\bar{r}_u+\frac{\sum_{k\in\bar{U}}s(u,k)(r_{k,i}-\bar{r}_k)}{\sum_{k\in\bar{U}}|s(u,k)|} \tag{5}$$

式中：$\bar{U}$ 表示与用户 $u$ 相似度较高的近邻集，$r_{k,i}$ 表示近邻集中第 $k$ 个用户对产品 $i$ 的评分，用户 $u$ 与近邻集中第 $k$ 个用户的相似性表示为 $s(u,k)$，表示用户 $u$ 对产品的平均评分。式(3)中取近邻集中近邻用户对产品 $g$ 评分的均值，作为目标用户对产品的预测评分；式(4)以用户之间的相似度作为权重对平均打分法进行改进；式(5)中不仅考虑到了权重，还考虑到了用户评分尺度与偏好不同的影响。

针对于用户-产品矩阵稀缺性，刘庆鹏等[15]提出了综合均值优化方法来弥补稀缺性带来的冷启动问题。该方法首先利用评分矩阵中的行与列估计评分矩阵中的未评分项，然后根据处理后的评分矩阵进行综合处理得到最终评分，从而提高了系统的推荐质量。

3）基于模型的推荐算法

上述 2 类算法直接根据评分矩阵寻找近邻并进行评分预测，主要适用于用户兴趣状况稳定的情况。在大型商务网站，面对大规模用户及大量产品，用户/产品间相似性的计算，特别是用户间相似性的计算，不仅耗时而且计算量大，在真实的商务环境中该类算法的性能优势不明显，因此，为了确保系统的高可扩展性，研究者提出了多种基于模型的推荐算法。该类算法应用统计学和机器学习算法对现有数据进行挖掘，根据现有数据推断并建立模型，运行时仅通过得到的模型进行评分预测，包括 Bayes 模型[16]、概率相关模型[17]、极

大熵模型[18]、基于聚类的 Gibbs 抽样算法[19]、基于马尔可夫决策过程模型、线性回归模型[11]等。

朴素 Bayes 分类模型假设样本的各个属性特征相互独立,将联合条件概率分布的计算分解独立的条件概率相乘,大大简化了计算量。但用户之间存在相互依赖性时,算法的准确性会大打折扣。Ungar 等[19]提出了一种聚类模型——Gibbs 抽样模型,该模型分别对用户和产品进行聚类,不仅能够改变用户/产品所在的类,而且能够同时改变含有该用户/产品的事件。模型包含 3 个参数:$P_k$、$P_l$、$P_{kl}$,其中 $P_k$ 表示随机选取的用户 $u_i$ 被分配到类 $k$ 中的概率;$P_l$ 表示随机选取的产品被分配到类 $l$ 中的概率;$P_{kl}$ 则表示 $k$ 类中的用户与 $l$ 类中的产品有关联(如用户喜欢/不喜欢该类中的产品)的概率。Gibbs 抽样需要在分配和参数估计两步骤之间不断迭代直到估计出的模型参数收敛。Sarwar 与 Karypis[11]考虑将线性回归模型用于预测用户评分。他们指出用余弦相似性与相关系数计算用户/产品间相似度时,如果在用户/产品空间 2 个评分向量之间的距离较远时,会导致较高的相似性,在这种情况下,根据用户-产品评分数据进行的预测其准确性会降低。该模型是在加权评分预测(见式(4))的基础上进行了改进,利用回归模型估计近邻用户 $u_k$ 对目标产品 $g$ 的评分,根据得到的估计值计算目标用户 $u_i$ 对产品 $g$ 的评分。该类算法的不足之处在于,模型建立之后需要根据用户兴趣的变化定期更新而模型的建立及更新过程需要耗费大量的计算。

## 1.2 基于内容的推荐算法

考虑到协同过滤算法主要关注用户-产品评分矩阵,忽略了用户信息(如年龄、性别、职业、地区等)和产品信息(如类型、规格、生产商等),基于内容的推荐算法主要解决如何根据用户和产品本身的特征进行合理推荐的问题[3]。算法通过提取用户/产品特征,学习用户兴趣模型,考察用户资料与候选推荐产品之间的匹配度,将匹配度最高的产品推荐给用户[20]。用户/产品特征的提取主要通过对用户/产品的文本描述为主。在信息获取中表征文本最常使用词频-逆文档频率法。该方法的定义如下:设有 $N$ 个文本文件,关键词 $k_i$ 在 $n_i$ 个文件中出现,将关键词 $k_i$ 在文件 $j$ 中出现的次数设为 $f_{ij}$,那么 $k_i$ 在 $j$ 中的词频 $T_{ij}$ 定义为:

$$T_{ij}=\frac{f_{ij}}{Z_{\max}f_{zj}} \tag{6}$$

式中:$Z$ 表示在文档 $j$ 中出现的关键词,分母的最大值可以通过计算 $j$ 中所有关键词的频率得到[21]。在许多文件中同时出现的关键词对于区分文件的关联性是没有贡献的[22]。因此,$T_{ij}$ 与这个关键词在文中出现的次的逆($I_i$)一起使用,

$$I_i=\log\frac{N}{n_i} \tag{7}$$

那么一个文件 $j$ 中的内容可以表示成向量 $\boldsymbol{d}_j=(w_{1j},w_{2j},\cdots,w_{ij})$,$w_{ij}$ 可以表示为

$$w_{ij}=\frac{f_{ij}}{Z_{\max}f_{zj}}\log\frac{N}{n_i} \tag{8}$$

该算法适用于用户及产品特征容易提取的情况,Fab 系统[23]就是一个典型的基于内容的推荐算法的应用。

## 1.3 基于二部图的推荐算法

Aggarwal 于 1999 年率先提出了基于二部图的推荐算法[24],该类算法仅关注用户是否选择某个产品,并不关心用户和产品是何种形式。在二部图算法中用户和产品被看作图的节点,如果用户选择某个产品,用户节点和产品节点之间就存在边,否则用户和产品节点之间不存在边。因此,通过用户与产品之间的选择关系建立用户-产品二部图模型,计算用户节点 $u_i$ 与未选择过的产品 $g_j$ 之间的相关性,根据相关性的大小向用户推荐其可能感兴趣的产品。假定用户个数为 $m$,产品个数为 $n$,那么 $m+n$ 个节点以及由于用户选择某个产品形成的边构成了如图 1 所示的二部图。

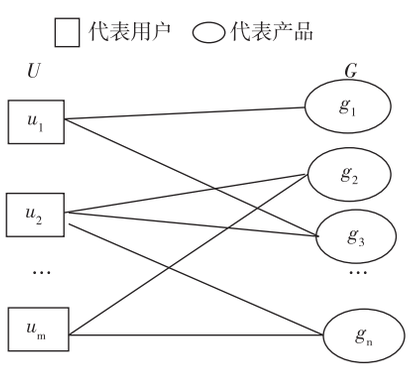


**图 1 基于二部图的网络结构图**

文献[25]中提出了基于资源分配的推荐算法。假设用 $U$ 代表用

户节点，用 $G$ 代表产品节点，则用户节点 $U$ 和产品节点 $G$ 以及节点之间因选择关系形成的边 $E$ 组成二部图网络结构。资源分配的过程通过如下步骤完成：①利用已知的用户-产品间的选择关系构建权重对象网络（如构建权重矩阵 $\boldsymbol{W}$）；②根据用户的历史数据确定用户的初始资源向量 $\boldsymbol{f}$；③根据初始资源向量与权重矩阵的乘积得到最终的资源分配；④根据最终的资源分配向用户推荐产品节点 $G$ 中资源分配较高且用户未选择过的产品。以第一个用户 $u_1$ 为目标，资源分配后，产品 $i$ 从产品 $j$ 中获取的资源值

$$\boldsymbol{W}_{ij}=\frac{1}{K_{\mathrm{j}}}\sum_{u=1}^{m}\frac{\partial_{\mathrm{ui}}\partial_{\mathrm{uj}}}{K_u} \tag{9}$$

式中：$K_{\mathrm{j}}$ 表示产品 j 被选择的次数，即产品 j 的度；$K_{\mathrm{u}}$ 表示用户 $u$ 的度。$\partial_{ui}$ 的定义

$$\partial_{\mathrm{ui}}=\begin{cases}1;\text{用户 } u \text{ 选择的产品 } i\\0;\text{用户 } u \text{ 未选择产品 } i\end{cases} \tag{10}$$

对于一个给定的目标用户，通过式(9)可以计算任意产品 $i$ 从产品 $j$ 中获得的资源，从而得到一个 $n$ 维的矩阵 $\boldsymbol{W}$，根据式(10) 可以得到一个 $n$ 维的 0/1 矢量，代表针对该个体的初始资源分配情况，将这个矢量记为 $\boldsymbol{f}$。最终的资源分配矢量可以表示为 $\boldsymbol{f}'=\boldsymbol{W}\times\boldsymbol{f}$。把目标用户没有看过的产品，按照 $\boldsymbol{f}'$ 中对应元素的大小进行排序 —— 值越大就说明用户偏好度越高，从而产生个性化推荐。

### 1.4　基于关联规则的推荐算法

算法的核心思想：关注用户的行为数据，从大量数据中抽取潜在有用的关联规则[26]，从而向用户推荐其感兴趣的产品。学者 Agrawal 和 Swami 最先提出了基于关联规则的算法，而最先成型的关联规则算法是经典的 Apriori 算法[27]。关联规则认为：假设 $I=\{i_1,i_2,\cdots,i_m\}$ 为项的集合，$D=\{t_1,t_2,\cdots,t_n\}$ 为交易数据库，该数据库中的每个事务 $t_i$ 均为非空子集并且每一个交易都有一个唯一的 TID(Transaction ID)与之相对应，对于 $I$ 中的子集 $X$，如果有 $X\subseteq Y$，那么称事务 $T$ 支持 $X$[28]。把形如 $X\Rightarrow Y$ 的蕴含式称作关联规则，其中 $X,Y\in I$ 且 $X\cup Y=\varnothing$，关联规则的先导和后继分别用 $X$ 和 $Y$ 表示。在关联规则 $X\Rightarrow Y$ 中交易数据库 $D$ 所包含 $X\cup Y$ 的百分比，即 $P(X\cup Y)$ 称作此关联规则的支持度；置信度是包含 $X$ 的事务中同时包含 $Y$ 的百分比，即条件概率 $P(Y|X)$[29]。支持度和置信度是关联规则算法中 2 个重要指标，如果既满足最小支持度的阀值又满足最小置信度的阀值，那么称该关联规则是有趣的。

### 1.5　基于社交网络的推荐算法

算法的核心思想：利用社交网络数据捕捉用户兴趣偏好及好友信息，并根据获取的数据为用户进行个性化产品推荐、好友推荐与信息流的会话推荐。

近年来，基于社交网络的推荐已经发展成为个性化推荐领域的研究热点之一。社交网络通过汇集不同领域、职业、地区、年龄的人员，极大地丰富和拓展了人们的交流圈，同时激发了局域社交网络营销中蕴藏的巨大商业价值与潜力[30]。相关领域的研究人员将基于社交网络的推荐分成两类：基于邻域的社会化推荐与基于网络结构的社会化推荐。

基于邻域的社会化推荐利用社交网络将用户的好友关系数据与用户历史行为及兴趣数据相结合，向目标用户推荐好友喜欢的产品集合。一般情况下，用户更倾向于选择自己熟悉的好友所推荐的产品，因此算法中需要考虑用户与好友之间的熟悉程度及兴趣相似程度，用户 $u$ 对产品 $i$ 的兴趣 $P_{ui}$ 可用公式(11)表示[31]：

$$P_{\mathrm{ui}}=\sum_{v\in f(u)}(f_{\mathrm{uv}}+s_{\mathrm{uv}})r_{\mathrm{vi}} \tag{11}$$

式中：$f(u)$ 表示用户 $u$ 的好友集；$f_{\mathrm{uv}}$ 表示用户 $u$ 与用户 $v$ 之间的熟悉程度；$s_{\mathrm{uv}}$ 表示用户 $u$ 和用户 $v$ 之间兴趣爱好的相似度；$r_{vi}$ 表示用户 $v$ 对产品 $i$ 的偏好（如果用户 $v$ 喜欢产品 $i$，$r_{\mathrm{vi}}=1$；否则 $r_{\mathrm{vi}}=0$）。

基于网络结构的社会化推荐分别以用户社交网络图、用户-产品二部图的形式来表示用户的社交网络及用户对产品的行为。通过获取的社交网络数据，将社交网络图和用户-物品二部图组合成一个网络图。该算法首先依据用户与好友之间的熟悉程度及兴趣相似度、用户对产品的偏好度对网络图中边的权重进行定义，然后计算用户节点与物品节点之间的相关性，最后按相关性的大小选取用户没有直接选择的产品并生成推

荐列表[31]。

## 2 各类推荐算法的优劣及其典型应用

上述各种推荐算法各有优劣,协同过滤的推荐算法优缺点都较明显,应用也最为广泛。基于内容的推荐算法通过分析产品的特征属性进行推荐,在文本信息推荐领域应用最为成熟,在对其他产品进行推荐时,易受特征提取技术的制约;基于二部图的推荐算法将用户和产品表示为二分图模型,根据模型为用户进行个性化推荐,但由于在计算过程中未考虑权重导致准确度降低,研究人员针对该问题提出了多种改进仍在不断探索;基于关联规则的推荐算法根据在用户数据中提取的关联规则进行推荐,在零售业领域应用最为成功,但在实际应用中,关联规则较难提取;基于社交网络的推荐受到了很多网站的重视,该类算法利用好友数据向目标推荐产品或好友,可以减轻"信息超载"现象,但在大型网站中获取用户好友数据存在困难。表 2 还给出了各类算法的典型应用系统。

**表 2 主流推荐算法优劣及典型应用表**

| 推荐算法 | 优点 | 缺点 | 典型应用 |
|---|---|---|---|
| 协同过滤 | 推荐性能随时间的推移不断提高;<br>能够向用户提供新的兴趣点;推荐个性化、自动化程度高;不需要领域知识;<br>能够处理复杂的非结构化对象 | 用户-产品矩阵的稀疏性;<br>可扩展性、冷启动问题;<br>对用户的评分数据依赖性大 | MovieLens、Netflix GroupLens、Amazon、当当、淘宝、CDNow、360buy、MovieFinder |
| 基于内容的推荐 | 结果直观,可解释性好;<br>不需要领域知识;<br>不需要用户评分数据; | 受新用户/新产品的限制;<br>推荐结果缺乏惊喜;<br>"度"对推荐算法产生不良影响 | Fab 系统 |
| 基于二部图的推荐 | 推荐内容不受限;<br>推荐结果具有多样性; | 受新用户/新产品的限制;<br>没有考虑用户评分差异量,推荐质量及个性化程度较低 | P2P 交流网 |
| 基于关联规则的推荐 | 易发现新的兴趣点;<br>不需要领域知识; | 关联规则难抽取、耗时;<br>个性化程度低 | ILOG |
| 基于社交网络的推荐 | 利用好友进行推荐增加了用户对推荐结果的信任度;<br>有利于推荐长尾商品 | 在实际系统中难以获取用户好友数据 | Clicker 视频推荐网站 |

## 3 性能评价指标

个性化推荐系统常采用的性能评价指标包括:平均绝对误差、均方根误差、标准平均误差、召回率、准确率。

1)平均绝对误差:用于衡量用户预测评分与实际评分之间的平均绝对误差,定义如式(12)所示

$$M=\frac{1}{n}\sum_{a=1}^{n}\mid p_{ia}-r_{ia}\mid \tag{12}$$

2)均方根误差定义如式(13)所示:$R_{\mathrm{m}}$ 表示均方根误差,

$$R_{\mathrm{m}}=\sqrt{\frac{1}{n}\sum_{a=1}^{n}\mid p_{ia}-r_{ia}\mid^{2}} \tag{13}$$

3)标准平均误差定义为

$$N=\frac{M}{r_{\max}-r_{\min}} \tag{14}$$

式中:$n$ 为用户 $i$ 已评过分的产品数量;$M$ 为平均绝对误差;$p_{ia}$ 与 $r_{ia}$ 分别为预测的用户评分和真实的用户评分;$R_{\mathrm{m}}$ 为均方根误差;$N$ 为标准平均误差;$n_i$ 为系统中所包含的用户-产品对;$r_{\max}$ 为用户评分的最大值;$r_{\min}$ 为用户评分的最小值[32]。

召回率表示推荐列表预测的用户喜欢的产品与系统中用户喜欢的所有产品的百分比。计算公式为

$$R=\frac{N_l}{N_r} \tag{15}$$

准确率：定义为推荐列表中用户喜欢的产品在所有被推荐的产品中所占的比率，计算公式为

$$P=\frac{N_l}{N_s} \tag{16}$$

式中：$R$ 为召回率；$P$ 为准确率；$N_l$ 为用户喜欢的产品被推荐的个数；$N_r$ 为系统中用户喜欢的产品；$N_s$ 为所有被推荐的产品。

在评价系统时，召回率和准确率必须结合使用才能够对算法的优劣作出评价。Pazzani M 等[33]将两者综合，提出了 $F$ 指标，计算方法为

$$F=\frac{2PR}{P+R} \tag{17}$$

周涛等[34]在文献中提出，在评估算法的准确性时，可以利用平均排队值(Ranking score)法。设 $L_i$(已经根据用户兴趣进行了排序)表示用户未选择的产品数量，如果用户 $i$ 与产品 $j$ 之间存在选择关系，同时产品 $j$ 在排序时被排在了 $R_{ij}$ 位置，那么$(i, j)$的相对位置为

$$r_{ij}=\frac{R_{ij}}{L_i} \tag{18}$$

此外，Pearson 关联[35]、Speaman 关联[36]和 Kendall's Tau[37]也可以作为评价系统准确性的指标。Pearson 关联定义为

$$C=\frac{\sum(x-\bar{x})(y-\bar{y})}{n\sqrt{\prod(x-\bar{x})^2}\sqrt{\prod(y-\bar{y})^2}} \tag{19}$$

式中：$n$ 为向量维度，$x$ 和 $y$ 表示用户向量与产品向量对应位置的评分。在排名相关性的计算方面，还可以借助 Kendall's Tau 方法，计算结果越大则预测越精准，定义为

$$T_a=\frac{C-D}{\sqrt{(C+D+T_R)(C+D+T_P)}} \tag{20}$$

式中：$C$ 为系统中预测正确的用户兴趣偏序数；$D$ 为预测错误的用户兴趣偏序数；$T_R$ 为用户实际评分相同的产品个数；$T_P$ 为预测值相同的产品个数。

距离标准化指标[23]、半衰期指标[38]、ROC 曲线[39]等指标也可以用来度量推荐系统的性能。推荐系统不仅需要高的准确性，关键要得到用户的认同。因此刘建国等[40]提出了除准确性之外的其他指标，包括推荐产品的流行性、多样性、覆盖率、新颖性及用户满意度等。

推荐系统自提出以来，工业界与学术界的相关研究者们不断探索，虽然已经提出了多种推荐算法，但对于哪种算法的性能最优目前没有仍未有统一的定论。数据集不同，算法的表现也会存在差异。Joonseok 等[41]对影响个性化推荐算法精准度的因素进行了分析，研究表明用户数量、产品数量以及评分矩阵的密集度会影响算法的精准度。例如，基于用户的协同过滤算法对产品的数量有很大的依赖性，而基于产品的协同过滤算法对用户数量有很大的依赖性。

## 4　结　论

推荐系统已经成为缓解“信息超载”问题的有利工具。与搜索引擎相比，推荐系统的优势在于能够主动收集用户的特征资料，挖掘蕴含在用户行为数据中的有效信息并定制性地向用户提供其可能感兴趣的产品或信息，同时通过及时跟踪用户的需求变化自动调整信息服务的方式和内容。目前推荐系统已经应用到多个领域，比如电子商务领域(如 Amazon. com、eBay 等)、网页标签领域(如 Fab、sesamr. com 等)、新闻领域(如 GroupLens 等)、电影领域(如 MovieLens，Netflix，Moviefinder. com 等)、音乐领域(如 Ringo 等)，其中在电子商务领域中的应用最为成熟。

虽然推荐系统已经在众多领域得到了研究与应用，但是随着系统规模的不断扩大以及用户与产品数量的指数级增长，用户对产品的评分数据变得更为稀疏。以 Movielens 数据集为例，该数据集为协同过滤算法

研究中使用最多的数据集之一,其中 Movielens 1M 数据集包含了包含6 039位用户对3 883部电影的1 000 209条评分记录,但该数据集的稀疏度达到了 95.73%,过高的稀疏度严重降低了推荐系统的性能;当新用户与新产品进入系统后,用户、产品信息的缺少使得推荐系统面临着冷启动问题,即无法准确向新用户推荐符合其兴趣偏好的产品。一些推荐算法需要提取用户/产品的特征,从文本信息中提取特征比较容易,但从多媒体信息(如视频、音频、图像等)中提取特征受到技术上的制约,造成推荐系统无法准确获取用户与产品的特征。此外,由于协同过滤算法需要在整个数据空间进行计算,在数据集较小的情况下,其推荐效果较好,但是面对上百万用户/产品时,该类算法的可扩展性不佳,降低了系统的时效性和精准性[42]。

为了缓解用户-产品评分矩阵的稀疏性、冷启动及可扩展性问题,文献[43]中提出了一种将用户聚类与产品聚类技术相结合的个性化推荐算法。该算法首先依据评分矩阵对用户进行聚类,通过计算目标用户与聚类中心的相似性进行评分预测,然后结合产品聚类技术产生推荐;文献[44-45]将矩阵分解技术引入到推荐系统中。此外,对现有算法进行改良与并行化运算已成为解决电子商务环境下数据矩阵稀疏性、可扩展性等问题的研究热点,不少学者对推荐系统的评价指标、多维度推荐等进行研究和扩展。

## 参考文献:


[1]ADOMAVICIUS G, TUZHILIN A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions[J]. IEEE Trans on Knowledge and Data Engineering, 2005,17(6):734-749.

[2]RESNICK P, VARIAN H R. Recommender system[J]. Communication of the ACM, 1997, 40(3):56-58.

[3]赵良辉,熊作贞. 电子商务推荐系统综述及发展研究[J]. 电子商务, 2013, 35(12):58-60.

[4]崔春生,吴祁宗,王莹.用于推荐系统聚类分析的用户兴趣度研究[J]. 计算机工程与应用,2011, 47(7):226-228.

[5]GOLDBERG D, NICOLS D. Using collaborative filtering to weave an information tapestry[J].Communications of the ACM, 1992,35(12):61-70.

[6]刘发升, 洪营. 基于用户特征属性和云模型的协同过滤推荐算法[J]. 计算机工程与科学, 2014, 36(6):1172-1176.

[7]许海玲. 互联网推荐系统比较研究[J]. 软件学报, 2009, 20(2):350-362.

[8]孙光福, 吴乐, 刘淇, 等 . 基于时序行为的协同过滤推荐算法[J]. 软件学报, 2013, 24(11):2721-2733.

[9]RESNICK P, IAKOVOU N, SUSHAK M, et al. GroupLens: an open architecture for collaborative filtering of netnews [C]// Proceeding of the 1994 Computer Supported Cooperative Work Conference.North Carolina:ACM, 1994:175-186.

[10]LINDEN G, SMITH B, YORK J. Recommendations tem-to-item collaborative filtering[J]. IEEE Internet Computing, 2003, 7(1):76-80.

[11]SARWAR B, KARYPIS G, KONSTAN J, et al. Item-based collaborative filtering recommendation algorithms[C]//International World Wide Web Conferences. Hongkong:ACM,2001:285-295.

[12]DESHOANDE M, KARYPIS G. Item-based top-n recommendation algorithms[J]. ACM Trans Information System, 2004, 22(1):143-177.

[13]张光卫,李德毅,李鹏,等 . 基于云模型的协同过滤推荐算法[J]. 软件学报, 2007, 18(10):2403-2411.

[14]黄创光, 印鉴, 汪静, 等 . 不确定近邻的协同过滤推荐算法[J]. 计算机学报, 2010,33(8):1369-1377.

[15]刘庆鹏, 陈明锐. 优化稀疏数据集提高协同过滤推荐系统质量的方法[J]. 计算机应用, 2012, 32(4):1082-1085.

[16]CHIEN Y H, GEORGE E I. A Bayesian model for collaborative filtering[C]// Proceeding of the Steventh International Workshop Artificial Intelligence and Statistics.Florida:[s. n.], 1999.

[17]GETOOR L, SAHAMI M. Using probabilistic relational models for collaborative filtering[C]// Proceeding of the Workshop Web Usage Analysis and User Profiling (WEB KDD'99). San Diego:[s. n.], 1999.

[18]PAVLOV D, PENNOCK D. A maximum entropy approach to collaborative filtering in dynamic, sparse, high-dimensional domains[C]// International Conference on Neural Information Processing.Cambridge:MIP Press,2002:1465-1472.

[19]UNGAR L H, FOSTER D P. Clustering methods for collaborative filtering[C]//Proceedings of the 1998 workshop on Recommen Dation Systems.Menlo Park:AAAI Press,1998:84-88.

[20]常璐. 高校图书馆 E-learning 支持服务研究[D]. 上海:东华大学, 2013.

[21]SALTON G. Automatic text processing: the transformation, analysis, and retrieval of information by computer[M]. Boston: Addison-Wesley, 1989.

[22]刘玲. 基于 Topsis 思想的内容推荐算法研究[J]. 数学的实践与认识, 2012, 42(16):113-119.
[23]BALABANOVIC M, SHOHAM Y. Fab: content-based collaborative recommendation[J]. Communications of the ACM, 1997, 40(3):66-72.
[24]蔡红蕾. 二部图网络结构算法在推荐系统中的应用[D]. 秦皇岛:燕山大学,2014.
[25]ZHOU T, JING L L, SU R Q, et al. Effect of initial configuration on network-based recommendation[J]. Europhys Lett, 2008,81(5):58004.
[26]肖波, 徐前方, 蔺志青, 等. 可信关联规则及其基于极大团的挖掘算法[J]. 软件学报, 2008, 19(10):2597-2610.
[27]PINTO H, HAN J, PEI J, et al. Multi-dimensional sequential pattern mining[C]// Conference on Information and Knavledge Management. Atlanta: ACM, 2001: 81-88.
[28]杨红菊, 梁吉业. 一种有效的关联规则的挖掘方法[J]. 计算机应用, 2004, 24(3):88-89.
[29]殷红, 许彦如, 王长波. 考虑信誉的网络交易可视化研究[J]. 东华大学学报(自然科学版), 2013, 39(4):514-518.
[30]黄仁, 孟婷婷. 个性化推荐算法综述[J]. 中小企业管理与科技(中旬刊), 2015(8):271-273.
[31]项亮. 推荐系统实战[M]. 北京:人民邮电出版社, 2012:151-152.
[32]王国霞, 刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用, 2012, 48(7):66-76.
[33]PAZZANI M, BILLSUS D. Learning and revising user profiles: The identification of interesting Web sites[J]. Machine Learning, 1997, 27(3):313-331.
[34]ZHOU T, REN J, MEDO M, et al. Bipartite network projection and personal recommendation[J]. Physical Review E, 2007, 76(4): 046115.
[35]RODGERS J L, NICEWANDER W A. Thirteen ways to look at the correlation coefficient[J]. The American Statistician, 2012, 42(1): 59-66.
[36]SPEARMAN C. The proof and measurement of association between two things[J]. American Journal of Psychology, 1904, 15(1): 72-101.
[37]KENDALL M. A new measure of rank correlation[J]. Biometrika, 1938, 30: 81-93.
[38]BREESE J, HECHERMAN D, KADIE C. Empirical analysis of predictive algorithms for collaborative filtering[C]// Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence.San Francisco:Morgan Kaufmann, 1998:43-52.
[39]SWETS J A. Information retrieval systems[J]. Science, 1963, 141(3577):245-250.
[40]JOONSEOK L, MINGXUAN S, GUY L. A comparative study of collaborative filtering algorithms[J/OL]. (2012-03-14)[2016-01-04].https://arxiv.org/abs/1205.3193.
[41]刘建国, 周涛, 郭强, 等. 个性化推荐系统评价方法综述[J]. 复杂系统与复杂性科学, 2009, 6(3):1-10.
[42]应毅,刘亚军,陈诚. 基于云计算的个性化推荐系统[J]. 计算机工程与应用,2015,51(13):111-117.
[43]GONG S. A collaborative filtering recommendation algorithm based on user clustering and item clustering[J]. Journal of Software, 2010, 5(7):745-752.
[44]涂丹丹,舒承椿,余海燕. 基于联合概率矩阵分解的上下文广告推荐算法[J]. 软件学报,2013,24(3):454-464.
[45]BAUER J, NANOPOULOS A. A framework for matrix factorization based on general distributions[C]// Proceedings of the 8-th ACM Conference on Recommender Systems. Silicon Valley: ACM Press,2014: 249-256.

(责任编辑:李艳)