

# 协同过滤推荐项目优化处理的初步研究

何光辉<sup>1</sup> 鲍丽山<sup>2</sup> 王蔚韬<sup>3</sup> 周 戈<sup>3</sup>

(重庆大学数理学院 重庆400044)<sup>1</sup> (江苏省电力调度通信中心 南京210024)<sup>2</sup>

(重庆大学计算机学院 重庆400044)<sup>3</sup>

**摘要** 协同过滤(CF)推荐系统应用知识发现技术为实时交易的用户提供个性化的产品或服务推荐。这些系统在电子商务领域取得了很大的成功。但是,在克服CF推荐系统的算法可伸缩性和推荐质量这两个根本性挑战方面还存在许多问题。本文分析了传统的CF算法,并介绍了一种提高推荐质量的新方法,我们称这种新方法为CF算法的推荐优化。从我们的分析可得,我们的方法相比传统的CF算法提供了更高的质量保证。

**关键词** 协同过滤推荐算法,伸缩性,推荐质量,推荐优化

## A Preliminary Study about Optimization of Recommendations in Collaborative Filtering Algorithms

HE Guang-Hui<sup>1</sup> BAO Li-Shan<sup>2</sup> WANG Wei-TAO<sup>3</sup> ZHOU Ge<sup>3</sup>

(School of Science, Chongqing University, Chongqing 400044)<sup>1</sup> (Communication Center of Jiangsu Electric Power Regulation, Nanjing 210024)<sup>2</sup>

(School of Computer, Chongqing University, Chongqing 400044)<sup>3</sup>

**Abstract** Collaborative Filtering(CF) Recommender systems apply knowledge discovery techniques to the problem of making personalized recommendations for products or services during a live interaction. These systems are achieving widespread success in the E-commerce or Web. However, there remain important research questions in overcoming two fundamental challenges for CF Recommender systems. They are the scalability of CF algorithms and the quality of recommendations. In this paper, we analyze the traditional CF algorithms, and introduce a novel approach to improve the quality of the recommendations for the users. We name it Optimization of Recommendations in CF Algorithms. From our analysis, it is obviously that our approach provides better quality than traditional CF algorithms.

**Keywords** Collaborative filtering recommendation algorithms, Scalability, Quality of recommendations, Optimization of recommendations

## 1 引言

推荐系统应用知识发现(Knowledge Discovery)技术为进行电子商务交易的用户提供个性化的信息、产品或服务。由于推荐系统的应用潜力巨大,针对推荐系统尤其是推荐算法的研究已经成为电子商务的一个活跃的研究领域。CF(Collaborative Filtering)类推荐算法目前在电子商务的研究和应用中都比较成功。但随着电子商务交易的频繁,越来越多的顾客参与进来,同时网站提供的项目也成千上万,传统的基于用户(user-based)的CF类推荐算法日益暴露了其弱点:1)CF算法缺乏伸缩性(Scalability);2)CF算法的推荐质量有待提高。

针对基于用户的CF算法的弱点,Badrul等提出了基于项目的CF类算法,从而提高了CF算法的可伸缩性,同时还能一定程度上提高推荐质量。但推荐算法的伸缩性和推荐质量仍然是研究人员非常感兴趣的研究领域。

虽然对CF类推荐算法的研究较多,但针对推荐算法(包括CF类推荐算法)的推荐项目进行优化处理的文献却非常少见。本文从另一个新的角度,通过对推荐项目进行优化处理,进而提高推荐质量。

## 2 相关工作

Tapestry<sup>[4]</sup>是最早的协同过滤推荐算法之一,该系统依赖于具有相同背景的人群的明确观点、兴趣、爱好,如一个办公室的同事。Tapestry算法的明显确定就在于要求顾客之间是相互认识的,而这在互联网上几乎是办不到的。其后,出现了几个自动排序的推荐算法,如GroupLens, Ringo and

Video推荐算法,这些算法不能针对具体用户给出个性化的服务,因而应用不是特别广泛。

随着统计技术和数据挖掘技术的加入,如贝叶斯网络(Bayesian networks)、聚类(Clustering)、劝告(Horting),使得推荐算法尤其是CF类推荐算法的性能有了很大的提高。贝叶斯网络通过训练数据集建立一个决策树模型(可以离线建立),然后在决策树中快速、准确地查找“邻居”(具有相似爱好、兴趣的顾客),贝叶斯网络适用于用户的兴趣、爱好变化不大的环境。

聚类技术将用户分为具有类似爱好的若干类。一旦类形成后,就可以从中获得单个用户将要购买的项目序列。聚类技术一般适用于人数较少的环境。但只要聚类形成,运行效果都非常好,因为一个类包括的用户数较少。其缺点是由于电子商务网站的顾客数量变化较大,因而需要不断实时修改聚类结果,这会降低推荐算法的效率。

劝告(Horting)技术是运用基于图理论的推荐算法,在图中结点代表用户,而两个用户间的类似程度用边表示。预测是通过沿着边寻找相近的点(用户)并综合考虑邻近用户的兴趣、爱好作出的。

以上这些算法都是基于用户(user-based)的推荐算法,这类算法的共同缺点就是可伸缩性相对较差以及推荐质量不高。针对这两个缺点,Badrul等提出了基于项目(item-based)的推荐算法。该方法思想的直观理解就是用户会对自己早期已经购买的项目的相似项目感兴趣,而对自己早期没有购买的类似项目不会感兴趣。因而只需要将项目之间的相似性计算出来(可以离线计算),就可以预测用户最有可能将要购买的N个项目的排序。当然该方法可以部分地解决可伸缩性问

题以及推荐质量问题。但该方法对早期购买项目较少的顾客来说推荐效果不是很好。推荐系统的主要步骤如图1所示。

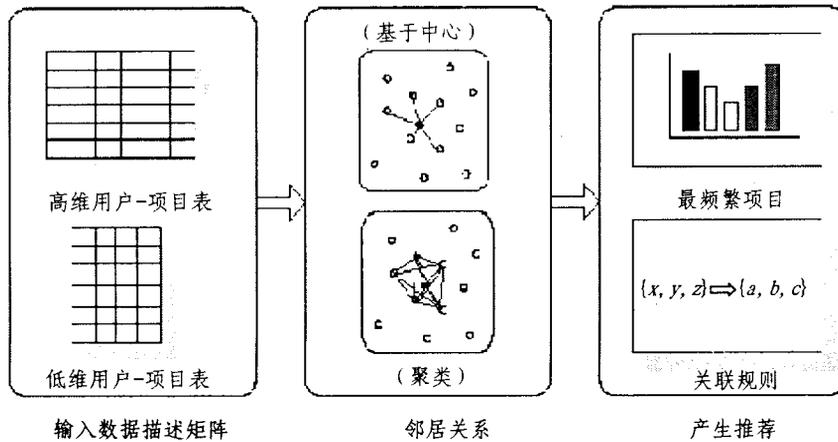


图1 推荐系统的三个主要步骤<sup>[11]</sup>

从以往的推荐算法的步骤可以看出在提出推荐项目后没有考虑推荐项目的优化问题,本文利用项目分类的方法来对推荐项目进行优化,从而达到提高推荐项目的推荐质量的目标。

### 3 CF 类推荐系统

推荐系统在电子商务中运用数据分析技术,如统计、数据挖掘、图论等,帮助用户查找他们最有可能想要购买的商品、服务序列。常用的方法是列出 Top-N 个推荐项目给用户。CF 类算法是到目前为止比较成功的一类推荐算法,其基本思想是根据给定用户  $u_s$  的已经购买的项目向用户  $u_s$  提供推荐或预测项目。

#### 3.1 CF 类算法的过程

推荐算法的主要目的就是向某一给定用户提出购买项目

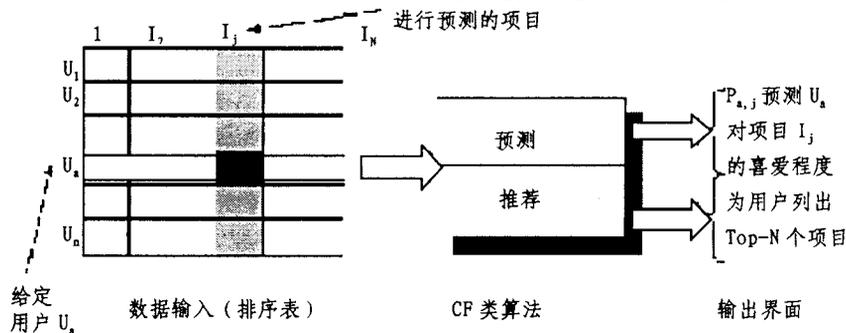


图2 CF 类推荐算法过程<sup>[1]</sup>

CF 类推荐算法的过程可以用图2来说明。CF 算法一般用  $m \times n$  阶用户-项目的喜爱程度(或购买数量),如果是0则表示用户对该项目没有评价或购买。

#### 3.2 CF 类算法的类型

根据对用户-项目矩阵  $A$  的不同研究方法,CF 类推荐算法可以分为以下两类:

1. 基于用户(user-based)的算法:该类算法根据全部的用户-项目数据产生用户-项目矩阵  $A$ ,根据需求采用降维技术后,利用统计、数据挖掘等技术找到给定顾客  $u_s$  的邻居,邻居关系确定后推荐系统可以利用统计技术对顾客  $u_s$  的邻居所购买的项目(亦即邻居的兴趣、爱好)进行排序,然后给顾客  $u_s$  提出推荐项目,这就是我们熟悉的“最近邻居”算法。

的建议,其推荐的根据是该顾客以往的兴趣、爱好或者类似顾客的兴趣、爱好。在典型的 CF 类算法中,我们通常设  $m$  个用户集合为  $U = \{u_1, u_2, \dots, u_m\}$ ,  $n$  个项目集合为  $I = \{i_1, i_2, \dots, i_n\}$ ,对每个用户  $u_s$  有已经购买的或评价的项目集合  $I_{u_s}$ ,集合  $I_{u_s}$  (常用数值表示)表明了顾客  $u_s$  对其已经购买项目的兴趣、爱好程度(或者是购买的数量)。由于  $I_{u_s} \subset I$ ,因此  $I_{u_s}$  可以为空集。对某一给定顾客  $u_s$ ,CF 类推荐算法的主要任务就是根据项目之间的相似性和邻居关系得出以下两种形式的结果:

·推荐 即提供一个  $N$  维列表  $I' \subset I$ ,向用户推荐他们最喜欢的  $N$  个项目。当然这  $N$  个项目是该用户以前没有购买的。即  $I' \cap I_{u_s} = \Phi$ ,通常我们称之为 Top-N 推荐方法。

·预测 设  $P_{i,j}$  表示用户  $u_s$  对项目  $i_j$  的喜爱程度(数值型),对所有项目进行计算后得到一个项目相似度矩阵。运用统计技术比较后得出一个顾客对项目的喜爱程度序列。

2. 基于项目(item-based)的算法:该类算法首先将用户对项目的兴趣、爱好建立一个模型,建立模型的过程一般采用机器学习的方法,如贝叶斯网络、聚类和关联规则方法。根据所建立的模型计算出项目之间的相似度,并且可以构造出相似度矩阵。然后根据顾客  $u_s$  已经购买的项目确定其最有可能将要购买的项目。

### 4 推荐项目的优化

#### 4.1 问题的提出

从以往的 CF 类推荐算法过程可以看出,基于用户或者基于项目的推荐算法过程主要分为三个步骤:

1. 基于用户的 CF 类推荐算法

- 5) 在客户和商家之间有一个协商过程;
- 6) 客户可以提供证书来说明她有权享有特殊的定价或待遇;
- 7) 客户只有在对商品付款后才能收到信息商品;
- 8) 在 NetBill 服务器进行一次交易前, 客户可能需要得到某个第四方的允许;
- 9) 通信的隐私性和完整性不会被参与交易的其他方所观察或修改。

## 2.2 NetBill 交易协议

本文使用符号“ $X \Rightarrow Y$ ”表示  $X$  发送指定的消息给  $Y$ 。基本的交易协议包含 8 个步骤:

- 1)  $C \Rightarrow M$  Price request
- 2)  $M \Rightarrow C$  Price quote
- 3)  $C \Rightarrow M$  Goods request
- 4)  $M \Rightarrow C$  Goods, encrypted with a key  $K$
- 5)  $C \Rightarrow M$  Signed Electronic Payment Order
- 6)  $M \Rightarrow N$  Endorsed EPO (including  $K$ )
- 7)  $N \Rightarrow M$  Signed result (including  $K$ )
- 8)  $M \Rightarrow C$  Signed result (including  $K$ )

2.2.1 价格协商 在 NetBill 交易协议中, 客户和商家之间的价格协商过程发生在协议的前两步。首先客户向商家发送一个价格请求, 请求中包含了客户的身份标识(可能是用户的假名)、客户对商品的出价以及交易号  $TID$  等。商家随后向用户返回标价, 这个消息中包含了产品号( $ProductID$ )、商品的价格和交易号  $TID$ 。其中, 交易号  $TID$  是一个可选项, 客户和商家可以使用这个交易号对商品的价格进行反复的协商直到双方满意为止。

2.2.2 商品发送 交易协议中的第 2 和第 3 步是商家发送电子商品的阶段。客户和商家对商品的价格达成一致后, 客户向商家发送商品请求。在接到客户的这个请求后, 商家向客户发送用商品密钥  $E_k$  (采用对称密钥机制) 加密了的电子商品以及由商家产生的一个唯一的电子定单序号  $EPOID$ 。

2.2.3 付款 客户收到商家发送的已加密的电子商品后, 客户就开始决定是否进行交易。如果客户不希望继续交易, 那么在这一步她可以安全地终止交易, 她的利益不会受到任何损害; 如果客户决定继续交易, 那么她将不能再终止协议的继续进行。

如果客户决定继续交易, 那么她将签名一个电子支付定单 EPO。EPO 包含有两个部分的内容, 其中一部分是公开的信息, 包括用户的身份标识(可能是假名)、商家的身份标识、产品号、价格、 $EPOID$ 、对收到的电子商品的密码学校验和以及对商品请求数据的密码学校验和等, 公开的部分商家和服务器都能够读取; 另一部分是加密的, 该部分包含客户的帐户信息和真实身份等内容, 只有 NetBill 服务器才能够读取。整个 EPO 的内容如下。

EPO 的公开部分包括: 1) 客户的身份标识(可能是假名); 2) 可读的产品号(从第 2 步得来的); 3) 协商的价格(从第 2 步得来的); 4) 商家的身份标识; 5) 已加密了的商品的密码学校验和; 6) 产品请求数据(从第 1 步得来的)的密码学校验和; 7) 客户的帐号以及一个帐户验证临时值的密码学校验和; 8) 全局唯一的  $EPOID$ 。

EPO 中加密的部分包括:

- 1) 证明客户真正身份的票据; 2) 任何需要的认证代币; 3) 客户的 NetBill 帐号; 4) 帐户验证临时值; 5) 客户的备注字段。
- 用户将构造好的 EPO 数字签名之后发送给商家。

商家收到客户的电子定单 EPO 后, 对定单中的公开部分

进行检查, 并确认  $EPOID$ 、产品号、商品的价格以及客户收到的电子商品的密码学校验和等信息是否与自身的记录相符。在这一步, 商家可以安全地终止交易, 他的利益也不会受到任何损害。如果商家决定继续交易, 那么他将不能再终止交易的继续进行。

如果商家决定继续交易, 商家将在客户的 EPO 后面附上商家的 NetBill 帐号、商家的注释以及商品密钥  $K$ , 随后商家对此消息进行数字签名并将交易提交给 NetBill 服务器。

NetBill 服务器收到商家提交的交易请求后, 将检查交易的序号  $EPOID$  是否有重复、交易客户的帐户余额是否足够等信息, 并决定是否进行交易。如果交易成功, 商家把交易的结果发送给商家, 这个结果将作为交易的收据。收据的内容如下:

$Result, Identity, Price, ProductID, M, K, EPOID$

其中  $Result$  为交易结果代码,  $Identity$  为客户的身份标识(可能为客户的假名),  $Price$  为商品的价格,  $ProductID$  为交易的商品号,  $M$  为商家的身份标识,  $K$  为商品密钥,  $EPOID$  为此次交易的电子定单号。

这个收据包含了交易过程中的有用信息, 它可作为商家和客户在争论中的凭证使用。商家在收到服务器的这个收据后, 将这个收据的一个拷贝转发给客户。如果客户没有收到商家发送的收据, 她可以直接向银行索取。

## 2.3 NetBill 协议的特点

NetBill 交易协议满足电子商务的原子性要求。电子商务中的原子性概念是由 Tygar 提出的<sup>[3]</sup>。Tygar 认为电子商务协议应该满足 3 类原子性的要求, 即钱原子性、商品原子性和确认发送原子性。

钱原子性(Money Atomicity): 钱既不能被创生, 也不能被销毁。客户支付的钱等于商家收到的钱。

商品原子性(Goods Atomicity): 客户只有付款后才能收到商品; 同理, 商家只有发送了商品后才能收到付款。

确认发送原子性(Certified Delivery): 客户能够确认商家所发送的商品正是她所订购的, 商家也能够证明客户所声称的商品正是商家所发送的。

交易协议满足原子性可以解决交易过程中的很多问题, 同时为交易结束后各方之间的争论提供了不可否认的证据。

NetBill 交易协议还提供了构造客户假名的机制, 这种机制能够保护客户的真实身份, 从而能够保护客户的隐私(如购物习惯和购买数量等信息)。另外, NetBill 交易协议还提供了一种使得客户能够证明自己在组中的成员关系的证书机制, 这使得 NetBill 交易协议能够支持商家的打折(例如, 若商家对客户采取会员制的管理策略, 那么对于不同等级的会员, 商家会提供商品不同程度的打折服务)。

NetBill 交易协议还具有其他一些特性, 本文就不再一一列举。

## 2.4 NetBill 交易协议存在的问题

不少研究者从安全性和原子性角度对 NetBill 协议进行了细致的分析<sup>[4-6, 13]</sup>。本文将从时限约束的角度分析 NetBill 交易协议中存在的一种缺陷。

下面举个简单的例子来说明 NetBill 电子交易协议中存在的一个问题。

假设商家  $M$  决定在时间期限  $\{T_1, T_2\}$  范围内对所有的购物客户实施打折的优惠。这时, 如果有某位客户  $C$  在接近于  $T_2$  的时刻  $T_1$  提出了购物请求, 那么商家会怎么处理呢? 这时商家有 3 种选择:

- 1) 接受购物请求并为客户实施打折优惠; 2) 拒不接受客