基于分布式数据的隐私保持协同过滤推荐研究

张 锋 常会友

(中山大学信息科学与技术学院 广州 510275)

摘 要 针对分布式数据存储结构的协同过滤推荐隐私保持问题,以可交换的密码系统为主要技术,设计了一个协议,集中解决其核心任务——在保持用户隐私前提下对项目评分.准确度与数据集中存放一样,但能保持各分站点下用户评分数据的隐私.基于安全多方计算理论和随机预言模型,证明了协议的安全性,分析了协议的时间复杂度和通信耗费.

关键词 推荐系统;协同过滤;隐私保持;安全多方计算;随机预言模型

中图法分类号 TP391

Research on Privacy-Preserving Collaborative Filtering Recommendation Based on Distributed Data

ZHANG Feng CHANG Hui-You

(*School of Information Science and Technology, Sun Yat-Sen University, Guangzhou* 510275)

Abstract Privacy-preserving data mining is a cutting-edge research direction in recent years. As one of its sub-directions, privacy-preserving collaborative filtering aims at protecting users' privacy while providing high-quality recommendations efficiently. To reserve privacy in collaborative filtering recommender systems under distributed data scenario, the core challenge——how to securely rate a specific item——is addressed. A protocol employing commutative encryption as its major privacy-preserving technique is introduced. This protocol produces the same results as the traditional memory-based collaborative filtering recommender systems while preventing any user' ratings from being known by other sites rather than by itself. Based on secure multi-party computation and random oracle model, the protocol's security is proved. The protocol's computation complexity and communication costs are analyzed as well.

Keywords recommender system; collaborative filtering; privacy preserving; secure multi-party computation; random oracle model

1 引 言

协同过滤推荐技术基于相似用户群的兴趣向目标用户产生推荐,是当前最成功、使用最广泛的推荐技术之一.除了提高可扩展性和改善推荐质量两大传统挑战[1,2],和其它数据挖掘技术一起,协同过滤推荐的隐私保持问题近年来引起了越来越多学者的研究兴趣.

数据挖掘中隐私保持是近年来学术界的一个研究热点,已在很多数据挖掘任务中取得了成果,比如关联规则[3,4]、决策树[5]、聚类[6]、奇异点探测[7]、

收稿日期:2006-04-04;修改稿收到日期:2006-06-05.本课题得到国家自然科学基金(60573159)、广东省自然科学基金重点项目基金(05100302)资助.张 锋,男,1974 年生,博士研究生,讲师,主要研究方向为数据挖掘中隐私保持、机器学习算法等.E-mail:zhfeng@mail.sysu.edu.cn.常会友,男,1962 年生,博士,教授,博士生导师,主要研究领域为协同软件研究、智能算法设计、复杂系统建模等.

Bayesian 网络[8]等. 本文的研究内容是协同过滤推荐中的隐私保持[9~12]. 协同过滤推荐的隐私保持最常使用的技术和其它数据挖掘隐私保持研究类似,大致可以分为两类:一类是基于加密的技术[3,4];另外一类是所谓的随机扰乱技术(Randomized Perturbation Techniques,RPT)[11]. 前一种主要应用于数据分布式存储的隐私保持数据挖掘研究;后一种虽然常见于数据集中式存储的情况,但也可以应用于分布式数据存储情况. 另外还有一种据我们所知、并未见应用于协同过滤推荐的隐私保持技术是随机响应技术(Randomized Response Techniques)[13],这种技术的核心思想与 RPT 有点类似.

Canny 在文献[9,10]中第一次提出了基于 P2P 结构的协同过滤推荐隐私保持问题. 文献[9]采用了 SVD 技术[14]和极大似然技术产生推荐,Canny 把协同过滤任务约化为用户数据评分向量的反复相加,所以数据的隐私保持可以采用同态加密技术完成;文献[10]的隐私保持技术虽然与前者一致,但它采用基于 EM 的因子分析技术产生推荐,据称提高了推荐质量. 两篇论文都属于基于模型的协同过滤推荐技术研究范畴. 但 SVD 技术会导致信息损失[14],类似的因子分析技术也存在同样的问题,而且虽然两篇论文均给出了算法,并且宣称是基于安全多方计算的实际应用,但并没有给出其安全性的证明.

文献[11]针对基于集中式数据的隐私保持协同过滤推荐设计了一个解决方案,主要采用了随机扰乱技术. 虽然这种技术宣称可以很好地从扰乱后的数据构建足够准确的模型,但会导致推荐质量的下降,另外 Kargupta 等对随机扰乱技术的安全性提出了强烈的质疑[15]. 文献[12]是类似于文献[11]的一个研究,但是基于 SVD 技术.

对于数据挖掘的隐私保持研究成果,虽然国外已经有比较多的文献报道,国内却比较鲜见,但基于双方或多方的安全计算(算法、协议)研究偶有报道[16~17].

本文主要贡献有两点:(1)在已有研究基础上,第一次把安全多方计算理论应用于基于分布式数据存储的隐私保持协同过滤推荐,设计了一个安全协议. 协议在保证准确地进行协同过滤评分的前提下,确保各分站点评分数据隐私(参见第 2 节的隐私定义);(2)给出了利用安全多方计算理论[18]和随机预言模型[19],形式化证明多方安全协议安全性的方法,并分析了协议的时间复杂度和通信耗费.

本文第 2 节给出问题定义;第 3 节简单介绍与本文密切相关的安全多方计算基本概念;第 4 节是协议的描述及其安全性证明;第 5 节分析协议的时间复杂度和通信耗费;最后是对文章的小结.

2 问题定义

协同过滤技术通常使用的是用户-项目评分数据集,它根据一定的量度标准在评分数据集中找出目标用户的"最近邻居",然后参考这些"最近邻居"的"意见",预测目标用户对项的评分,从而根据预测评分产生推荐.

常见的最近邻居度量标准包括余弦相似性、相关相似性和修正的余弦相似性.

我们采用相关相似性度量,如公式

$$\mathrm{sim}(u,v)=\frac{\sum_{i\in\phi(u,v)}(R_{u,i}-\bar{R}_u)(R_{v,i}-\bar{R}_v)}{\sqrt{\sum_{i\in\phi(u,v)}(R_{u,i}-\bar{R}_u)^2}\sqrt{\sum_{i\in\phi(u,v)}(R_{v,i}-\bar{R}_v)^2}} \tag{1}$$

$\bar{R}_u$和$\bar{R}_v$分别表示用户 u 和用户 v 对已评分项目评分的算术平均值;$\phi(u,v)$是用户 u,v 共同评分项目集;$R_{u,i}$,$R_{v,i}$分别是用户 u、用户 v 对项目 i 的评分.

采用相似度由高至低计算得到目标用户 u 的最近邻居集 NBS_u后,用户 u 对项 i 的预测评分 $P_{u,i}$可通过式(2)计算:

$$P_{u,i}=\bar{R}_u+\frac{\sum_{v\in NBS_u}\mathrm{sim}(u,v)\times(R_{v,i}-\bar{R}_v)}{\sum_{v\in NBS_u}\mathrm{sim}(u,v)} \tag{2}$$

其中符号的含义同式(1). 推荐的候选项目一般是最近邻居中目标用户未评分项目. 为节约篇幅和更有效地说明本质问题,本文针对目标用户对指定项目的评分进行预测,这个协同过滤推荐核心问题提出隐私保持的计算模型. 一般的计算模型是对所有合适的候选项进行评分预测,而由对一项的评分推广到对多项的评分是容易的.

设有 $N(N\geq 3)$个站点,$(S_0,S_1,\cdots,S_{N-1})$,站点 i 的评分数据格式如式(3)所示.

$$\boldsymbol{S}_i=\begin{bmatrix}(v^i)_{1,1} & (v^i)_{1,2} & (v^i)_{1,3} & \cdots & (v^i)_{1,m}\\ (v^i)_{2,1} & (v^i)_{2,2} & (v^i)_{2,3} & \cdots & (v^i)_{2,m}\\ (v^i)_{3,1} & (v^i)_{3,2} & (v^i)_{3,3} & \cdots & (v^i)_{3,m}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ (v^i)_{n_i,1} & (v^i)_{n_i,2} & (v^i)_{n_i,3} & \cdots & (v^i)_{n_i,m}\end{bmatrix} \tag{3}$$

其中每个$(v_i)_{s,t}(1\leq s\leq n_i,\ 1\leq t\leq m)$表示用户 s 对第 t 个项的评分值.

共有 n_i 个用户：$\boldsymbol{S}_i=[\boldsymbol{U}_1\ \ \boldsymbol{U}_2\ \ \boldsymbol{U}_3\ \ \cdots\ \ \boldsymbol{U}_{n_i}]^{\mathrm{T}}$. 其中 $\boldsymbol{U}_k(1\leq k\leq n_i)$是用户评分向量(行向量)，$\boldsymbol{U}_k=((v^i)_{k,1},(v^i)_{k,2},(v^i)_{k,3},\cdots,(v^i)_{k,m})$.

每个用户用 m 个项描述：

$$\boldsymbol{S}_i=[\boldsymbol{I}_1\ \ \boldsymbol{I}_2\ \ \boldsymbol{I}_3\ \ \cdots\ \ \boldsymbol{I}_m],$$

其中 $\boldsymbol{I}_j(1\leq j\leq m)$是项目评分向量(列向量)

$$\boldsymbol{I}_j=((v^i)_{1,j},(v^i)_{2,j},\cdots,(v^i)_{n_i,j})^{\mathrm{T}}.$$

给定目标用户，$\boldsymbol{\mu}=(v_1,v_2,\cdots,v_m)$，使用协同过滤推荐技术，基于这 N 个站点的所有用户评分向量，准确地对某一指定项(目标项)τ 进行评分，同时希望数据隐私能最大限度得到保持：各个分站点的评分数据不被任何其它站点看到；虽然一些分站点最后能看到目标用户和所有全局最近邻居的相似度，但除了自己产生的最近邻居与目标用户的相似度外，并不能辨别出其它的相似度是属于哪个站点的哪个最近邻居的.

3　安全多方计算

本文设计的协议基于安全多方计算理论，其基本目标是在一些合理的假设基础上，在保证各参与方评分数据隐私的前提下计算指定项目的评分. 下面给出一些基本概念.

定义 1(计算不可区分).　两个使用 S 做索引的集合，$X\overset{\text{def}}{=}\{X_w\}_{w\in S}$和 $Y\overset{\text{def}}{=}\{Y_w\}_{w\in S}$，当且仅当：对每一个多项式长度的电路簇$\{C_n\}_{n\in N}$，每一个正多项式$p(\cdot)$和足够长的 $w\in S$，均满足

$$|Pr[C_n(X_w)=1]-Pr[C_n(Y_w)=1]|<\frac{1}{p(|w|)}.$$

我们称它们是计算不可区分的，记为 $X\overset{c}{\equiv}Y$.

实际上，我们考虑的可证安全性是建立在参与方是半诚实(semi-honest)参与方前提下的. 所谓半诚实参与方，简单地说，就是参与方遵守协议的要求，但是它可能会试图根据中间计算结果，获取一些额外信息. 为简单起见，下面给出安全两方计算定义，而不是多方，容易从安全两方计算扩展为安全多方计算，详情可参见文献[18].

定义 2(半诚实约束下的隐私性).　有函数 $f:\{0,1\}^*\times\{0,1\}^*\to\{0,1\}^*\times\{0,1\}^*$，其中 $f_1(x,y)$，$f_2(x,y)$分别表示 $f(x,y)$的第一和第二个元素，Π 是计算 f 的一个双方协议. 对输入(x,y)执行 Π 后，第一和第二部分的视图分别写成 $VIEW_1^{\Pi}(x,y)$和 $VIEW_2^{\Pi}(x,y)$，也就是$(x,r,m_1,\cdots,m_i)$和$(y,r,m_1,\cdots,m_t)$，其中 r 是相应方抛掷硬币结果而 m_i 是接收到的第 i 条消息. 如果存在多项式时间的算法 S_1 和 S_2，满足：

$$\{S_1(x,f_1(x,y),f_2(x,y))\}_{x,y\in\{0,1\}^*}\overset{c}{\equiv}\{VIEW_1^{\Pi}(x,y),OUTPUT_2^{\Pi}(x,y)\}_{x,y\in\{0,1\}^*}\text{ 和}$$

$$\{S_2(y,f_2(x,y),f_1(x,y))\}_{x,y\in\{0,1\}^*}\overset{c}{\equiv}\{VIEW_2^{\Pi}(x,y),OUTPUT_1^{\Pi}(x,y)\}_{x,y\in\{0,1\}^*},$$

那么，Π 秘密地计算了函数 f.

Π 秘密地计算 f，当且仅当 Π 在半诚实模型下是安全的，详细证明见文献[18].

基于定义 2，在半诚实模型下，如果一个协议安全地计算了 f，那么，每个半诚实参与方在参与了协议计算后获取的所有信息也同样能够通过参与方的输入和输出在多项式时间内获得. 也就是说，如果每一方都可以多项式的时间通过该方的输入和输出在多项式时间内模拟出与真实视图计算不可分的模拟视图，那么这个协议就是安全的，详细证明可参考文献[18].

4　隐私保持的协同过滤协议

要解决第 2 节中提出的问题，可以采取类似文献[9,10]中的方法，但其缺点在第 1 节中已有论述；另一种比较明显的方法是，各分站点的最近邻居集匿名传送，但各分站点仍然能看到评分数据原始值，而且通信量也比较大. 我们采用这样的计算模型：目标用户的评分向量通过发起方提交到各个分站点，每个分站点根据第 2 节提到的协同过滤技术计算目标用户的本地最近邻居集以及相应相似度；利用可交换的加密系统，把这些分布于各分站点的相似度安全地合并于发起方，经解密后，求得全局最近邻居的相似度下限值，把这个值向各个分站点广播，各分站中相似度超过这个值的最近邻居组成全局最近邻居的本地子集，然后基于这些全局最近邻居的本地子集，各分站点在保证第 2 节数据隐私定义的前提下，协同计算，求得对某一指定项的评分，进而产生推荐. 注意到，这个计算模型的推荐准确度和数据集中存放的一样.

4.1　可交换的加密系统

要满足上述的安全计算定义，我们主要采用可交换的加密系统，参考文献[3,21]，我们得到定义 3.

定义 3(可交换的加密系统).　一个可交换的

加密系统 $F=(M,K,f,g)$，其中，M 为消息集；$K=(E,D)$ 为加密/解密对集，E,D 分别是公钥和密钥集；$f:E\times M\rightarrow M$ 为定义在有限域内的多项式时间加密函数；$g:D\times M\rightarrow M$ 为多项式时间解密函数. $f_e(x)\equiv f(e,x)$，$g_d(x)\equiv g(d,x)$，表达式 $a\in_r A$ 表示 a 按照均匀分布从 A 随机选取.

(1) 对一切 $e_1,e_2,\cdots,e_n\in E$；$d_1,d_2,\cdots,d_n\in D$；$m\in M$；$(i_1,i_2,\cdots,i_n)$，$(j_1,j_2,\cdots,j_n)$，$(s_1,s_2,\cdots,s_n)$ 和 $(t_1,t_2,\cdots,t_n)$ 是 $(1,2,\cdots,n)$ 的任意 4 个随机排列，满足：

$f_{i_1}(f_{i_2}(\cdots(f_{i_n}(m))\cdots))=f_{j_1}(f_{j_2}(\cdots(f_{j_n}(m))\cdots))=T$ 和 $g_{s_1}(g_{s_2}(\cdots(g_{s_n}(T))\cdots))=g_{t_1}(g_{t_2}(\cdots(g_{t_n}(T))\cdots))$.

(2) 对一切 $e_1,e_2,\cdots,e_n\in E$，对任意 $m_1,m_2\in M$，如果 $m_1\neq m_2$，有足够大的 k，满足

$$Pr(f_{e_1}(f_{e_2}(\cdots(f_{e_n}(m_1))\cdots))=f_{e_1}(f_{e_2}(\cdots(f_{e_n}(m_2))\cdots)))<\frac{1}{2^k}.$$

(3) 对 $x,y,z\in_r M$，$e\in_r E$，$\langle x,f_e(x),y,f_e(y)\rangle$ 的分布和 $\langle x,f_e(x),y,z\rangle$ 的分布是计算不可分的.

第一个属性是说，加密的结果与加密顺序无关；第二个属性则说明两个明文不同的话，加密后的密文不可能相同；第三个属性是说给定明文 x 及其密文 $f_e(x)$（但不泄漏 e），对任意新明文 y，攻击者不能在多项式时间内区分密文 $f_e(y)$ 和从密文域内均匀随机选取的数 z，因此不能在多项式时间内加密 y 或解密 $f_e(y)$.

4.2 协 议

设有 $N(N\geq 3)$ 个站点 $0,1,\cdots,N-1$，不失一般性，设发起协同过滤计算的站点是 0，最近邻居数最多为 NN，邻居相似度值域 S（离散量值域），目标用户 μ，被评分项 τ，每个站点都有符合随机预言模型[19]的针对 S 的 Hash 函数 h（参见 4.3 节），有针对 $h(S)$ 的可交换密码系统 $F=(M,K,f,g)$.

首先借助图 2 和图 3 来说明可交换加密的核心思想. 这种加密核心思想已应用于如文献[3,4]的隐私保持关联规则挖掘，本文将其应用于协同过滤评分预测.

图 2 和图 3 分别显示了两轮加密传送循环后的示意图，其它轮次的省略，它们演示了下面协议步 4 的的运行过程. 这样经过 $N-1$ 轮的传送后，各分站点产生的数据（相似度）均经过所有分站点加密. 解密过程类似，但由于数据集已经合并到一个站点，所以每轮不是每个分站点都参与解密和发送数据，只是当前拥有数据的分站点参与该操作.

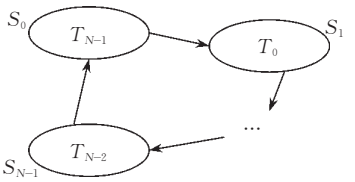

图 2 第一轮加密循环后各分站点结果示意图

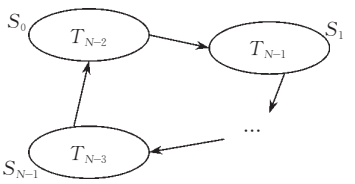

图 3 第二轮加密循环后各分站点结果示意图

协议具体步骤详述如下：

1. 站点 0 把接收到的目标用户 μ 的评分向量发向其它每个站点.

2. //每个站点分别求本地最近邻居及其相似度

对 $i=0,1,\cdots,N-1$ 的每个站点，

计算目标用户 μ 在站点 i 的最近邻居集 M_μ^i，M_μ^i 中每个邻居均已对 τ 评分，$NBS_\mu=\sum_{i=0}^{N-1}M_\mu^i$，$M_\mu^i$ 的第 j 个最近邻居表示为

$(userid_j,score_j^i,(I_{j_1}^i,score_{j_1}^i),(I_{j_2}^i,score_{j_2}^i),\cdots,(I_{j_t}^i,score_{j_t}^i))$. //其中的 $(I_{j_k}^i,score_{j_k}^i)$，$1\leq k\leq t$ 是邻居 $userid_j$ 的
//(已评分项，评分值) 对；$score_j^i$ 是目标用户 μ 和
//邻居 $userid_j$ 的相似度；M_μ^i 中所有邻居按照相
//似度由高到低排列

3. //加密各分站点本地最近邻居的相似度

对 $i=0,1,\cdots,N-1$ 每个站点

{

产生 (e^i,d^i)，$e^i\in_r E,d^i\in_r D$，

$T_i=f_{e^i}(h(score_1^i))$，

对站点 i 的 $j=2,\cdots,NN$ 各个邻居

{

第 j 个最近邻居相似度 $score_j^i$ 如果与 $score_{j-1}^i$ 不同，则对其施加 Hash 操作，使得 $u_j^i=h(score_j^i)\in M$，令 $S_j=f_{e^i}(u_j^i)$，$T_i=T_i\cup S_j$.

}

}

4. //作循环加密，到这一步结束，每个站点 i 都保留了
//站点 $(i+1)$ mod N 产生并执行了 Hash 操作，且
//被每个分站点加密的最近邻居集

For $r=0,1,\cdots,N-2$ //作 $N-1$ 轮循环

{

If $(r==0)$ Then

//第一轮，已经在上一步加密了，直接发到下个站点

每个站点 i 把 T_i 发送到相邻的下一个站点 $(i+1 \bmod N)$

Else

每个站点 i

//各站点加密接收到的数据，并

//往相邻下一站点发送

{

$k=(i-r) \bmod N$，

$T=$空集，

对每个 $t\in T_k$，

$\{t=f_{e^i}(t), T=T\cup t\}$

$T_k=T$，

把 T_k 送至与 i 相邻的下一个站点$(i+1 \bmod N)$.

}

}

对每个站点 i，

//最后一轮加密操作后不用传送

{

$k=(i+1-N) \bmod N$；

$T=$空集；

对每个 $t\in T_k\{t=f_{e^i}(t), T=T\cup t\}$

$T_k=T$；

}

5. //先把各分站点数据发到站点 1，最后合并到站点 //0，理由参考 4.3 节证明

对 $i=2,3,\cdots,N-1$ 的每个站点

把 $T_{(i+1) \bmod N}$ 发送到站点 1；

站点 1：令 $F=\bigcup_{i=1}^{n-1} T_{(i+1) \bmod N}$，把 F 里面的元素作随机排列，把 F 发送给站点 0；

站点 0：$F=F\cup T_1$，把 F 里面的元素作随机排列；

6. 把站点 0 中的 F 发到站点 1；

//合集经除站点 0 外各个站点解密

对 $i=1,2,\cdots,N-1$ 的各站点

{

$T=$空集；

对每个 $t\in F$

$\{t=g_{d^i}(t), T=T\cup t\}$

$F=T$；

把 F 发送到下一相邻站点$(i+1) \bmod N$

}

7. //合集在站点 0 完全解密得到原始值

在站点 0

{

$T=$空集；

对每个 $t\in F$

$\{t=g_{d^0}(t), T=T\cup t\}$

$F=T$；

对每个 $s\in S$

用 s 替换 F 中的每个 $h(s)$ 值；

对 F 中所有数据按相似度从高到低排序，选出第 NN 个数据对应的相似度 ν，记录下限为 ν 的数据个数 NN'（NN' 有可能大于 NN），为实际的最近邻居个数，把 ν 向各分站点广播；

}

8. 对 $i=0,1,2,\cdots,N-1$ 的每个站点

根据值 ν，计算本地站点中的满足相似度 $\geq v$ 的用户，组成“全局最近邻居”在本地 i 的分量，记为 $NBS_{\mu i}$，注意到全局最近邻居集 $NBS_\mu=\sum_{i=0}^{N-1} NBS_{\mu i}$，$NN'=|NBS_\mu|$；

9. 在站点 $i=0,1,\cdots,N-1$

计算 $(\mu, nu_\mu^i, de_\mu^i)$，其中 nu_μ^i 和 de_μ^i 是 $P_{\mu,\tau}=\bar{R}_\mu+\frac{\sum_{v\in NBS_\mu}\mathrm{sim}(\mu,v)\times(R_{v,\tau}-\bar{R}_v)}{\sum_{v\in NBS_\mu}\mathrm{sim}(\mu,v)}$ 的 $\frac{\sum_{v\in NBS_\mu}\mathrm{sim}(\mu,v)\times(R_{v,\tau}-\bar{R}_v)}{\sum_{v\in NBS_\mu}\mathrm{sim}(\mu,v)}$ 部分的分子分母在站点 i 的分量；

$nu_\mu^i=\sum_{v\in NBS_{\mu i}}\mathrm{sim}(\mu,v)\times(R_{v,\tau}-\bar{R}_v)$；

$de_\mu^i=\sum_{v\in NBS_{\mu i}}\mathrm{sim}(\mu,v)$；

10. 设 $|\mathrm{sim}(\mu,v)\times(R_{v,\tau}-\bar{R}_v)|\leq d_1$，$v\in NBS_\mu$，$|\sum_{v\in NBS_\mu}\mathrm{sim}(\mu,v)\times(R_{v,\tau}-\bar{R}_v)|\leq D_1=|NBS_\mu|\times d_1$，$|\mathrm{sim}(\mu,v)|\leq d_2$，$v\in NBS_\mu$，$|\sum_{v\in NBS_\mu}\mathrm{sim}(\mu,v)|\leq D_2=|NBS_\mu|\times d_2$；

在站点 0，从离散型均匀分布 $(0,2\times D_1]$（步长 $interval$）中产生随机数 d，从离散型均匀分布 $(0,2\times D_2]$（步长 $interval'$）中产生随机数 r，计算

$(nu_\mu^0+|NBS_{\mu 0}|\times d_1+d) \bmod 2\times D_1=\Sigma_0$；

$(de_\mu^0+|NBS_{\mu 0}|\times d_2+r) \bmod 2\times D_2=\Sigma_0'$，传给站点 1；

站点 1 计算

$(\Sigma_0+nu_\mu^1+|NBS_{\mu 1}|\times d_1) \bmod 2\times D_1=\Sigma_1$；

$(\Sigma_0'+de_\mu^1+|NBS_{\mu 1}|\times d_2) \bmod 2\times D_2=\Sigma_1'$，把结果传给站点 2，重复类似计算和传递，…，一直传到站点 $N-1$，结果为

$\left(\sum_{i=0}^{N-1} nu_\mu^i+|NBS_\mu|\times d_1+d\right)\bmod 2\times D_1$；

$\left(\sum_{i=0}^{N-1} de_\mu^i+|NBS_\mu|\times d_2+r\right)\bmod 2\times D_2$，把这两个值传给站点 0，令

$nu_\mu=\left(\sum_{i=0}^{N-1} nu_\mu^i+|NBS_\mu|\times d_1+d\right)\bmod 2\times D_1-d$；

$de_\mu=\left(\sum_{i=0}^{N-1} de_\mu^i+|NBS_\mu|\times d_2+r\right)\bmod 2\times D_2-r$；

如果 $nu_\mu<0$，令 $nu_\mu=2\times D_1+nu_\mu$；

如果 $de_\mu<0$，令 $de_\mu=2\times D_2+de_\mu$；

$nu_\mu=nu_\mu-D_1$，$de_\mu=de_\mu-D_2$；

在站点 0，用公式 $P_{\mu,\tau}=\bar{R}_\mu+\frac{nu_\mu}{de_\mu}$ 计算目标用户 μ 对各

τ 的评分.

4.3 安全性证明

由文献[21]的引理 2,容易得定理 1.

定理 1. 整数 n,分布

$$\begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ f_e(x_1) & f_e(x_2) & \cdots & f_e(x_n) \end{pmatrix}$$ 和分布 $$\begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ z_1 & z_2 & \cdots & z_n \end{pmatrix}$$ 是计算不可分的,其中, $1\leq i\leq n$,可交换的密码系统 $F, x_i, z_i\in_r M, e\in_r E$.

证明. 略.

协议的安全性还依赖于随机预言模型(random oracle model)[19]. 在随机预言模型下,Hash 函数 $h: V\rightarrow M$ 是理想的,也就是说 $h(v)$ 是理想化地被随机 oracle 计算的:每次一个新的 $v\in V$,都有一个独立的随机 oracle 找到 $x\in_r M$,满足 $x=h(v)$. 在 ROM 假设下,Hash 值不会冲突而且看起来是随机的.

定理 2. 在半诚实模型假设中,在随机预言模型和安全多方计算理论条件下,协议满足第 2 节中的安全性定义.

证明. 协议只有在步 4~7、步 10 发生了数据交换. 我们需要证明在安全多方计算理论和 ROM 假设下,5 个步骤中,根据输入输出多项式时间内计算得到的模拟视图和真实视图是计算不可分的,协议就是安全的. 步 10 的安全性,实质上是安全地求数值和问题,我们的协议采用文献[22]提到的安全求和方法. 对每一方来说,除了本地输入和全局输出,还可能"泄露"一些信息,关键在于如果可证根据本地输入和全局输出多项式时间内得到的模拟视图,和真实视图是计算不可分的,那么协议就是安全的,这是安全多方计算理论的核心观点,也在文献[18]中得到了严格证明.

在步 4,第一轮循环后,站点 i 看到的是集合 $T_{(i-1) \bmod N}$ 中所有数据,但由于已经过 Hash 和加密操作,所以能看到的额外信息是 $|T_{(i-1) \bmod N}|$.

设 $p=(i-1) \bmod N$, $q=|T_{(i-1) \bmod N}|$.

设每个站点 i 收到的站点 $(i-1) \bmod N$ 发过来的所有相似度组成集合 $S^p=(s_1^p, s_2^p, \cdots, s_q^p)\subseteq S$. 设有分布

$$R=\begin{pmatrix} u_1^p & u_2^p & \cdots & u_q^p \\ f_{e^p}(u_1^p) & f_{e^p}(u_2^p) & \cdots & f_{e^p}(u_q^p) \end{pmatrix},$$

$$E=\begin{pmatrix} u_1^p & u_2^p & \cdots & u_q^p \\ y_1^p & y_2^p & \cdots & y_q^p \end{pmatrix}.$$

$s_j^p\in S^p$, $u_j^p=h(s_j^p)$, $y_j^p\in_r M$,其中,$1\leq j\leq q$; R 和 E 分别是协议第一轮循环后站点 i 的真实视图和模拟视图,由定理 1,R 和 E 是计算不可分的.

其它轮次循环的安全性类似得证.

步 5 安全性证明方法和步 4 有不同的地方,主要是因为步 4 中分站点的数据是各不相同的,但在步 5 中由于数据合并,所以数据可能有重复. 真实视图的重复数据需要在模拟视图中准确地表示出来,才能保证两个视图是计算不可分的. 因为站点 0 和站点 1 的情况类似,所以只给出站点 0 的证明过程.

合并后共有 $W=NN\times N$ 个数据,去掉重复数据后共 R 个.

设站点 0 收到的所有相似度原始值,剔除了重复值后为 $s=(s_1, s_2, \cdots, s_R)\subseteq S$. 令 $I=(u_1\ \ u_2\ \ \cdots\ \ u_R)$,其中 $u_j=h(s_j)$, $1\leq j\leq R$.

站点 1 的真实视图是类似下面的分布:

$$R=\begin{pmatrix} u_{i_1} & u_{i_2} & \cdots & u_{i_w} \\ f_e(u_{i_1}) & f_e(u_{i_2}) & \cdots & f_e(u_{i_w}) \end{pmatrix}.$$

注意到为简单起见,我们用 f_e 代表多次加密操作.

$I'=(u_{i_1}\ \ u_{i_2}\ \ \cdots\ \ u_{i_w})\supseteq I$,其中 $i_k\in(1,2,\cdots,R)$, $1\leq k\leq w$, I' 可能比 I 多了些重复元素. 令 $IDX=(i_1, i_2, \cdots, i_w)$,显然 $IDX\supseteq(1,2,\cdots,R)$.

令 $X=\begin{pmatrix} u_1 & u_2 & \cdots & u_R \\ o_1 & o_2 & \cdots & o_R \end{pmatrix}$,构造函数

$$Q(X, IDX)=\begin{pmatrix} u_{i_1} & u_{i_2} & \cdots & u_{i_w} \\ o_{i_1} & o_{i_2} & \cdots & o_{i_w} \end{pmatrix}.$$

令 $IX=\begin{pmatrix} u_1 & u_2 & \cdots & u_R \\ f_e(u_1) & f_e(u_2) & \cdots & f_e(u_R) \end{pmatrix}$,

$VX=\begin{pmatrix} u_1 & u_2 & \cdots & u_R \\ y_1 & y_2 & \cdots & y_R \end{pmatrix}$,其中 $y_j\in_r M$.

把 IX 和 VX 分别代入函数 Q 中变量 X,分别得到真实视图和模拟视图. 由定理 1,IX 和 VX 是计算不可分的,而 IDX 可以多项式时间内计算得到(时间复杂度 $O(W\times R)\leq O(W^2)$),而函数 Q 是多项式时间可算的,所以 $Q(IX, IDX)$ 和 $Q(VX, IDX)$ 计算不可分,也就是真实视图和模拟视图是计算不可分的.

以上给出了由输入和输出模拟与真实视图计算不可分的模拟视图的过程,这个过程是多项式时间的.

另外需要解释的是为什么数据要经站点 1 然后集中到站点 0,而不是直接发到站点 0:如果所有数据直接发送到站点 0 的话,站点 0 能辨认出 Hash 操作和加密操作后的自己产生的数据,同时也获取了其它分站点产生的经 Hash 和加密后的数据,这

样一经比较，它就能区分出哪个站点的哪些数据跟自己是相同的了，这个“额外信息”是没有办法根据输入输出多项式时间内模拟的，所以会破坏协议的安全性定义. 但经过上述的特殊传输处理，这个问题就不存在了(注意到站点 1 产生的数据放在站点 0，而且数据传到 0 之前已作随机排列).

对步 6，在 $i=1,2,\cdots,N-1$ 的各分站点，是一个对数据逐步解密的过程，但所有数据始终至少被一个站点加密过，证明方法可以参考步 5.

步 7 的安全性证明方法跟上面的有点不同. 在这一步，所有数据都到达了站点 0，站点 0 完成所有步 7 的操作后，能看到所有站点的原始数据. 但由于数据到达站点 0 之前，都经过随机排列打乱了原来的顺序，站点除了能分辨出自己站点的相似度数据外，虽然也能看到所有其它站点的最近邻居相似度，但并不能分辨出这些相似度属于哪个分站点和属于哪个最近邻居，所以这一轮的数据传递仍满足我们的安全性定义，具体证明如下：

由于共有 N 个站点，设站点 $0,1,\cdots,N-1$ 产生的数分别为 $M_0,M_1,\cdots,M_{N-1}$ 个，站点 0 的实际视图是这 $\sum_{i=0}^{N-1}M_i$ 个数据的一个随机排列 V_r，而其输入是 $\sum_{i=0}^{N-1}M_i$ 个数. 现在我们需要的是根据这个输入/输出得到合适模拟视图.

首先要计算这 $\sum_{i=0}^{N-1}M_i$ 个数据分成 N 组(每组分别是 $M_0,M_1,\cdots,M_{N-1}$ 个数)的全排列. 实际上，由于站点 0 能够辨别出属于自己的 M_0 个数据，所以只是 $\sum_{i=1}^{N-1}M_i$ 个数据分成 $N-1$ 组(每组分别是 $M_1,M_2,\cdots,M_{N-1}$ 个数)的全排列，记为 PM，其大小为 $P(M)$.

模拟视图是 $\sum_{i=1}^{N-1}M_i$ 个数据的一个随机排列 S_r，对 $p\in_r PM$，$Pr(VIEW^{\text{real}}=p)=Pr(V_r=p)=1/P(M)=Pr(S_r=p)=Pr(VIEW^{\text{simulate}}=p)$；而 $p\notin_r PM$ 时，$Pr(VIEW^{\text{real}}=p)=Pr(V_r=p)=0=Pr(S_r=p)=Pr(VIEW^{\text{simulate}}=p)$.

可见模拟视图和实际视图是统计不可分的，根据文献[18]的统计不可分肯定是计算不可分定理，它们是计算不可分的，所以步 7 是安全的.

步 10 也发生了数据的传递. 实质上是各站点 i 的 $nu_\mu^i=\sum_{v\in NBS_{\mu i}}\text{sim}(\mu,v)\times(R_{v,\tau}-\bar{R}_v)$ 和 $de_\mu^i=\sum_{v\in NBS_{\mu i}}\text{sim}(\mu,v)$ 循环安全相加的过程，发起站点 0 发送这两个数给站点 1 的时候，已经分别加上一个随机数，这“掩饰”了原始数据值，站点 1 无法从接收到的经过处理后的数据看出原始数据值.

求和式中的 $\sum_{k=0}^{i-1}|NBS_{\mu k}|\times d_2$ 和 $\sum_{k=0}^{i-1}|NBS_{\mu k}|\times d_1$ 分量的作用是确保每个分站点的数据分量和为正数.

站点 i，模拟视图是分别从离散型均匀分布 $(0,2\times D_1]$ 和 $(0,2\times D_2]$ 随机选取的两个数组成的一个二元组 S_r，那么，对 $x\in_r(0,2\times D_1]$，$y\in_r(0,2\times D_2]$，有

$$\Pr(VIEW_i^{\text{real}}=(x,y))=$$

$$\Pr\Big(\Big(\sum_{k=0}^{i-1}nu_\mu^k+\sum_{k=0}^{i-1}|NBS_{\mu k}|\times d_1+d\Big)\bmod 2\times D_1=x,\ \Big(\sum_{k=0}^{i-1}de_\mu^k+\sum_{k=0}^{i-1}|NBS_{\mu k}|\times d_2+d\Big)\bmod 2\times D_1=y\Big)\Big)=$$

$$\frac{1}{2\times D_1\times interval}\times\frac{1}{2\times D_2\times interval'}=$$

$$\Pr(S_r=(x,y))=\Pr(VIEW_i^{\text{simulator}}=(x,y))$$

如果 x 和 y 至少有一个不属于两个离散型分布，类似等式同样成立. 所以计算不可分的视图可以在多项式时间内被模拟出来.

综合以上，由文献[18]的合成定理，协议是满足安全性定义的. 但要注意的是在这一轮证明中，如果 $P(M)=1$，也就是来自站点 0 之外的 $N-1$ 个站点的最近邻居相似度数据相同的情况下，站点 0 能够“辨认”出所有其它子站点的原始数据(和站点 0 的相同)，这时候协议是不能保证隐私的，但这种情况比较罕见.　　证毕.

5　时间复杂度和通信耗费分析

协议的时间复杂度和通信耗费如表 1 所示. 其中：

$|U|$ 是目标用户 μ 的用户—项向量长度；

N' 是每个分站点的用户评分向量数；

t 是一个评分项值编码位数；

C_h 是 Hash 操作的时间耗费；

C_e 是加密/解密操作的时间耗费；

t_1 是加密评分值的编码位数；

t_2 是评分值 v 的编码位数；

t_3 是各预测评分值在分站点分量的编码位数.

可以认为 $C_e \gg C_h$，$C_e \gg \log(NN \times N)$，$C_e \gg N'\log N' > N'$，$C_e \gg D_1$，$C_e \gg D_2$，因此协议的时间复杂度为 $O(2 \times C_e \times NN \times N + C_e \times N^2)$，因为在有些步骤分站点的处理可以同时进行，所以时间复杂度仍可以降低；而通信耗费为 $O(2 \times N^2 \times NN \times t_1)$.

表 1 各步骤时间复杂度和空间耗费

	时间复杂度	通信耗费
1	常量	$O(N \times \|U\| \times t)$
2	$O(N \times (N' + N'\log N'))$	无
3	$O((C_h + C_e) \times NN \times N)$	无
4	$O(C_e \times N^2)$	$O(N^2 \times NN \times t_1)$
5	$O(2 \times NN \times N)$	$O(N \times NN \times t_1)$
6	$O(C_e \times N \times NN)$	$O(N^2 \times NN \times t_1)$
7	$O((C_h + C_e) \times NN \times N + NN \times N\log(NN \times N))$	$O(N \times t_2)$
8	$O(NN \times N)$	无
9	$O(NN \times N)$	无
10	$O(2 \times D_1 + 2 \times D_2)$	$O(2 \times N \times t_3)$

5 结束语

本文提出的协议基于多方计算是安全的，但参与方必须大于 2，如果只有两个参与方的话，每个站点会知道其它站点的最近邻居相似度及对目标项的评分，并不满足我们的安全性定义. 而且参与方也必须满足半诚实约束，否则参与方可以“伪造”自身的输入来“获取”其它参与方的输入. 但基于半诚实约束的协议也是有广泛应用性的[18]，另外也可以参照文献[23]提出的增加一个输入承诺协议来保证参与方“半诚实”性质，限于篇幅，不再讨论.

参 考 文 献

1 Sarwar B., Karypis G., Konstan J., Riedl J.. Analysis of recommendation algorithms for E-Commerce. In: Proceedings of the 2nd ACM Conference on Electronic Commerce, Minneapolis, MN, USA, 2000, 158～167

2 Chee S. H. S., Han J. W., Wang K.. RecTree: An efficient collaborative filtering method. In: Proceedings of the 3rd International Conference on Data Warehousing and Knowledge Discovery, Munich, Germany, 2001, 141～151

3 Kantarcioglu M., Clifton C.. Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(9): 1026～1037

4 Vaidya J., Clifton C.. Secure set intersection cardinality with application to association rule mining. Journal of Computer Security, 2005, 13(4): 593～622

5 Agrawal R., Srikant R.. Privacy-preserving data mining. In: Proceedings of the 2000 ACM SIGMOD Conference on Management of Data, Dallas, Texas, USA, 2000, 439～450

6 Vaidya J., Clifton C.. Privacy-preserving k-means clustering over vertically partitioned data. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, D. C., USA, 2003, 206～215

7 Vaidya J., Clifton C.. Privacy-preserving outlier detection. In: Proceedings of the 4th IEEE International Conference on Data Mining, Brighton, UK, 2004, 233～240

8 Wright R., Yang Z.. Privacy-preserving Bayesian network structure computation on distributed heterogeneous data. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, 2004, 713～718

9 Canny J.. Collaborative filtering with privacy. In: Proceedings of the 2002 IEEE Symposium on Security and Privacy, Berkeley, California, USA, 2002, 45～57

10 Canny J.. Collaborative filtering with privacy via factor analysis. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, 2002, 238～245

11 Polat H., Du W.. Privacy-preserving collaborative filtering using randomized perturbation techniques. In: Proceedings of the 3rd IEEE International Conference on Data Mining, Melbourne, Florida, USA, 2003, 625～628

12 Polat H., Du W.. SVD-based Collaborative filtering with privacy. In: Proceedings of the 20th ACM Symposium on Applied Computing, Track on E-commerce Technologies, Santa Fe, New Mexico, USA, 2005, 791～795

13 Du W., Zhan Z.. Using randomized response techniques for privacy-preserving data mining. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, D. C., USA, 2003, 505～510

14 Aggarwal C. C. On the effects of dimensionality reduction on high dimensional similarity search. In: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Santa Barbara, California, USA, 2001, 256～266

15 Kargupta H., Datta S., Wang Q., Sivakumar K.. Random data perturbation techniques and privacy preserving data mining. Knowledge and Information Systems Journal, 2005, 7(4): 387～414

16 Qin J., Zhang Z. F., Feng D. G., Li B.. A protocol of comparing information without leaking. Journal of Software, 2004, 15(3): 421～427(in Chinese)
(秦 静，张振峰，冯登国，李 宝. 无信息泄露的比较协议. 软件学报，2004，15(3)：421～427)

17 Luo W. J., Li X.. The secure multi-party protocol of matrix product and its application. Chinese Journal of Computers, 2005, 28(7): 1230～1235(in Chinese)
(罗文俊，李 祥. 多方安全矩阵乘积协议及应用. 计算机学

报，2005，28(7)：1230～1235)

18 Goldreich O.. Foundations of Cryptography：Volume 2，Basic Applications. Beijing：Publishing House of Electronics Industry (Originated from Cambridge University Press)，2005

19 Bellare M.，Rogaway P.. Random oracles are practical：A paradigm for designing efficient protocols. In：ACM Conference on Computer and Communications Security，NY，USA，1993，62～73

20 Boneh D.. The decision diffie-hellman problem. In：Proceedings of the 3rd Algorithmic Number Theory Symposium，Portland，Orgeon，USA，1998，48～63

21 Agrawal R.，Evfimievski A.，Srikant R.. Information sharing across private databases. In：Proceedings of the ACM SIGMOD International Conference on Management of Data，San Diego，California，2003，86～97

22 Clifton C.，Kantarcioglu M.，Lin X. D.，Zhu M. Y.. Tools for privacy preserving distributed data mining. SIGKDD Explorations，2003，4(2)：28～34

23 Cachin C.. Efficient private bidding and auctions with an oblivious third party. In：Proceedings of the 6th ACM Conference on Computer and Communications Security，Kent Ridge Digital Labs，Singapore，1999，120～127

ZHANG Feng，born in 1974，Ph. D. candidate，lecturer. His research interests include privacy-preserving data mining，machine learning algorithms etc.

CHANG Hui-You，born in 1962，Ph. D.，professor，Ph. D. supervisor. His research interests include collaborative software research，intelligent algorithm design，complicated system modeling etc.

Background

This paper focuses on privacy-preserving data mining (PPDM)，an active research area in recent years. Currently，lots of PPDM results have been achieved in many data mining challenges，such as association rule mining，decision tree，clustering，outlier detection，Bayesian networks etc. Cryptography and random perturbation are two mostly used techniques. PPDM is still in its laboratory stage and there is still much work to do to put it into practice. Furthermore，we believe that there must be some new techniques to be figured out to do PPDM more effectively and more efficiently.

Based on previous results，the authors for the first time apply secure multi-party computation theory to the privacy-preserving collaborative filtering recommendation under distributed data scenario. The authors design a secure protocol to make collaborative filtering ratings be accurately computed while guaranteeing every involved party's privacy. The protocol employs commutative encryption systems as its major privacy-preserving technique. This protocol produces the same results as the traditional memory-based collaborative filtering recommender systems while preventing any user' evaluations from being known by other sites rather than by itself. Based on secure multi-party computation and random oracle model，the protocol's security is proved. The protocol's computation and communication costs are analyzed as well.

The work are supported by a grant from a General Program of National Natural Science Foundation of China (No. 60573159) and a grant from a Key Program of Guangdong Natural Science Foundation (No. 05100302). Roughly speaking，they both aim at resolving major challenges in collaborative computing and collaborative software development. Collaborative computing，as well as data mining，is a very large research direction，in which there is lots of work to do. Up to now，the research team has achieved some significant results in collaborative computing and data mining，including how to use soft computation techniques to support collaborative computation and designing algorithms and protocols to solve some key challenges in collaborative filtering recommenders.