两方参与的隐私保护协同过滤推荐研究

张 锋[1,2],孙雪冬[1],常会友[1],赵淦森[1]
(1. 中山大学软件学院,广东广州 510275;2. 中山大学广东省信息安全重点实验室,广东广州 510275)

摘 要: 隐私保护的协同过滤推荐研究致力于在确保高质、高效地产生推荐的同时有效地保护参与方的隐私. 在数据分布存储,参与方大于 2 的情形,已有研究针对其核心任务——对指定项进行评分预测,以可交换的密码系统为主要技术,设计了一个隐私保护计算协议. 但该协议不适用于参与方是 2 的情形. 以安全比较计算和安全点积计算为基础安全设施,设计了一个协议,解决参与方是 2 的情况下对指定项进行评分预测的隐私保护问题,从而解决了隐私保护的两方协同计算问题. 预测准确度与数据集中存放一样,证明了协议的正确性,并基于安全多方计算理论和模拟范例,证明其安全性,分析了时间复杂度和通信耗费.

关键词: 隐私保护数据挖掘; 安全多方计算; 推荐系统; 协同过滤
中图分类号: TP391 **文献标识码**: A **文章编号**: 0372-2112 (2009) 01-0084-06

Research on Privacy-Preserving Two-Party Collaborative Filtering Recommendation

ZHANG Feng[1,2], SUN Xue-dong[1], CHANG Hui-you[1], ZHAO Gan-sen[1]
(1. *Software School, Sun Yat-sen University, Guangzhou, Guangdong* 510275, *China*;
2. *Guangdong Key Laboratory of Information Security Technology, Sun Yat-sen University, Guangzhou, Guangdong* 510275, *China*)

Abstract: Privacy-preserving collaborative filtering aims at protecting participating parties' privacy while providing high-quality recommendations efficiently. In the case of the number of the participating parties is greater than 2, a protocol, employing commutative encryption as its major privacy-preserving technique, has been devised to address the issue of rating a specific item in scenarios with distributed data storage, which is a key challenge in privacy-preserving collaborative filtering recommendation in that scenarios. However, the protocol does not work when the number of the participating parties is exactly 2. Employing secure comparison and secure dot product as its fundamental security infrastructure, we design a privacy-preserving two-party collaborative computing protocol to address the challenge. This protocol produces the same results as the traditional memory-based collaborative filtering recommender systems. Based on secure multi-party computation theory and simulation paradigm, the protocol's security is proved. The protocol's computation complexity and communication cost are examined as well.

Key words: privacy preserving data mining; secure multi-party computation; recommender system; collaborative filtering

1 引言

隐私保护的数据挖掘研究权衡数据隐私和强大的数据挖掘功能,致力于在保证数据隐私的基础上进行准确、高效的数据挖掘工作,是近年来数据挖掘领域一个前沿研究方向,在很多数据挖掘研究领域取得了成果,如关联规则、决策树、聚类、离群点探测、Bayesian 网络等[2~8]. 也有研究针对典型的隐私保护数据挖掘问题开发或总结了一些非常有意义的基础工具或基础协议[9,10].

本文的研究内容是协同过滤推荐中的隐私保护[11~14]. 协同过滤推荐的隐私保护最常使用的技术和其它数据挖掘隐私保护研究类似,大致可以分为两类:一类是基于加密的技术[4,5];另外一类是随机扰乱技术[13]. 前者主要应用于数据分布式存储的隐私保护数据挖掘研究;后者常见于数据集中式存储的情况. 隐私保护数据挖掘技术还包括基于启发式方法[3],随机响应技术[15],k-匿名数据发布[16,17],等等.

John Canny 在文献[11]和[12]中第一次提出了基于 P2P 结构的协同过滤推荐隐私保护问题. 文献[11]采用了 SVD 技术[18]和极大似然技术产生推荐,Canny 把协同过滤任务约化为用户评分向量的反复相加,使得数据的

收稿日期:2007-08-20;修回日期:2008-10-27
基金项目:广东省自然科学基金重点项目(No. 05100302);广东省信息安全技术重点实验室开放基金;中山大学青年教师科研启动基金(No. 1131014)

隐私保护可以采用同态加密技术完成;文献[12]的隐私保护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. SVD 技术会导致信息损失[18],类似的因子分析技术也存在同样的问题. 虽然两篇论文均属于基于安全多方计算的实际应用,但并没有给出其安全性的严格证明.

文献[13]针对基于集中式数据,主要采用随机扰乱技术,设计了一个的隐私保护协同过滤推荐方案. 虽然这种技术宣称可以很好地基于扰乱后的数据对模型进行准确的重构,但会导致推荐质量的下降,另外 Kargupta H. 等对随机扰乱技术的安全性提出了强烈的质疑[19]. 文献[14]是类似文献[13]的一个研究,但是基于 SVD 技术.

文献[1]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护协同过滤推荐,设计了一个安全多方协同计算协议. 本文延续文献[1]的工作,主要贡献有两点:1) 针对文献[1]中的协议对参与方是 2 的情况下不能保持隐私的问题,采用新的工具和技术(安全比较计算和安全点积计算),设计了一个安全两方计算协议,解决参与方是 2 的情况下对指定项进行评分预测的隐私保护问题. 预测准确度与数据集中存放情形的一样;2) 论文证明了协议的正确性,并基于安全两方计算理论和模拟范例,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.

2　问题定义

协同过滤技术通常使用的是用户—项评分数据集,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",基于这些"最近邻居"的评分,计算预测目标用户对项的评分,进而产生推荐.

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(1):

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

$\bar{R}_u$ 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.

相似度值由高至低排列得到目标用户 u 的最近邻居集 NBS_u,用户 u 对项 i 的预测评分 $P_{u,i}$ 可通过公式(2)计算:

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

其中符号的含义同公式(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. 本文仅针对目标用户对指定项目评分进行预测这个协同过滤推荐核心问题提出隐私保护的计算模型. 一般的计算模型是对所有合适的候选项进行评分预测,容易由对一项的评分直接推广到对多项的评分.

设有 $N=2$ 个分站点:(S_0,S_1),站点 $S_i(i=0,1)$ 的评分数据格式如图 1.

每个 $(v^i)_{s,t}(1\le s\le n_i, 1\le t\le m)$ 表示用户 s 对第 t 个项的评分值.

共有 n_i 个用户:$S_i=[U_1,U_2,U_3,\cdots,U_{n_i}]^{\mathrm{T}}$. 其中 U_k $(1\le k\le n_i)$ 是用户评分向量(行向量),$U_k=((v_i)_{k,1},(v_i)_{k,2},(v_i)_{k,3},\cdots,(v_i)_{k,m})$.

每个用户用 m 个项描述:$S=[I_1,I_2,I_3,\cdots,I_m]$.

其中 $I_j(1\le j\le m)$ 是项目评分向量(列向量) $I_j=((v^i)_{1,j},(v^i)_{2,j},\cdots,(v^i)_{n_i,j})^{\mathrm{T}}$.

$$S_i=\begin{bmatrix}(v^i)_{1,1} & (v^i)_{1,2} & (v^i)_{1,3} & \cdots & (v^i)_{1,m}\\ (v^i)_{2,1} & (v^i)_{2,2} & (v^i)_{2,3} & \cdots & (v^i)_{2,m}\\ (v^i)_{3,1} & (v^i)_{3,2} & (v^i)_{3,3} & \cdots & (v^i)_{3,m}\\ & \cdots & & & \\ (v^i)_{n_i,1} & (v^i)_{n_i,2} & (v^i)_{n_i,3} & \cdots & (v^i)_{n_i,m}\end{bmatrix}$$

图 1　分站点 S_i 的评分数据

给定目标用户,$\mu=(v_1,v_2,\cdots,v_m)$,使用协同过滤推荐技术,基于这两个分站点的所有用户评分向量,找出全局最近邻居集,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:任一参与方除了获得对指定项的评分,不能准确地获得另一方任何数据(目标用户的评分向量 μ 除外).

3　协议

输入:两个分站点分别是 S_0、S_1;两个分站点的评分向量;每个分站点最近邻居数目 N;目标用户为 μ(用评分向量表示),目标项 τ.

输出:目标用户 μ 对目标项 τ 的评分.

不失一般性,设发起方为 S_0.

(1) 分站点 S_0 把接收到的目标用户 μ 的评分向量发向分站点 S_1.

(2) 分站点 S_0 和 S_1 分别计算目标用户 μ 在自身的最近邻居集合,M^0_μ 和 M^1_μ. M^0_μ 和 M^1_μ 中每个邻居均已对 τ 评分. $M^i_\mu(i=0,1)$ 的第 j 个最近邻居表示如下:

$(userid_j, score^i_j, (I^i_{j_1}, score^i_{j_1}), (I^i_{j_2}, score^i_{j_2}), \cdots, (I^i_{j_t}, score^i_{j_t}))$.

M^i_μ 中所有邻居按照相似度 $(score^i_j)$ 由高到低排列. S_0 和 S_1 的最近邻居相似度分别组成集合 T_0 和 T_1,$|T_0|=|T_1|=N$.

/ *其中的$(I^i_{j_k}, score^i_{j_k})$,$1 \le k \le t$ 是邻居 $userid_j$ 的(已评分项,评分值)对; $score^i_j$ 是目标用户 μ 和邻居 $userid_j$ 的相似度 */

(3) 设 2^k 是大于或等于 N 的最小的 2 的 k 次幂. 在分站点 S_0,往 T_0 末尾填入 $(2^k - N)$ 个 $-\infty$,使得 $|T_0| = 2^k$,这 2^k 个数按顺序存放在长度为 2^k 的数组 A 中;在分站点 S_1,往 T_1 头填入 $(2^k - N)$ 个 $+\infty$,使得 $|T_1| = 2^k$,这 2^k 个数按顺序存放在长度为 2^k 的数组 B 中.

(4)

①在 S_0,令 $i_0 = 0$, $j_0 = 2^k - 1$, $stop_0 = -1$;在 S_1,令 $i_1 = 0$, $j_1 = 2^k - 1$, $stop_1 = -1$.

②While$(i_0 != j_0)$

{

在分站点 S_0,计算 $m_0 = \lfloor \frac{i_0 + j_0}{2} \rfloor$;

在分站点 S_1,计算 $m_1 = \lfloor \frac{i_1 + j_1}{2} \rfloor$;

安全地比较 $A[m_0]$ 和 $B[m_1]$,如果 $A[m_0] < B[m_1]$,输出 1;否则输出 0.

分站点 S_0,如果得到比较结果为 0,则令 $stop_0 = m_0$, $i_0 = m_0 + 1$;如果得到比较结果为 1,则令 $j_0 = m_0$.

分站点 S_1,如果得到比较结果为 0,则令 $j_1 = m_1$;如果得到比较结果为 1,则令 $stop_1 = m_1$, $i_1 = m_1 + 1$.

}

安全地比较 $A[i_0]$ 和 $B[i_1]$,如果 $A[i_0] < B[i_1]$,输出 1;否则输出 0.

分站点 S_0,如果得到比较结果为 0,则令 $stop_0 = i_0$.

If $stop_0 != -1$,评分大于或等于 $A[stop_0]$ 的邻居组成全局最近邻居在分站点 S_0 的分量 NBS^0_μ,否则 NBS^0_μ 为空;

分站点 S_1,如果得到比较结果为 1,则令 $stop_1 = i_1$.

If $stop_1 != -1$,评分大于或等于 $B[stop_1]$ 的邻居组成全局最近邻居在分站点 S_1 的分量 NBS^1_μ,否则 NBS^1_μ 为空.

(5) 在分站点 S_0,计算 $nu^0_\mu = \sum_{v \in NBS^0_\mu} sim(\mu, v) \times (R_{v,\tau} - \bar{R}_v)$ 和 $de^0_\mu = \sum_{v \in NBS^0_\mu} sim(u, v)$;在分站点 S_1,计算 $nu^1_\mu = \sum_{v \in NBS^1_\mu} sim(\mu, v) \times (R_{v,\tau} - \bar{R}_v)$ 和 $de^1_\mu = \sum_{v \in NBS^1_\mu} sim(u, v)$.

在分站点 S_1,从域 $[a, b]$ 中随机均匀产生 r_1、r_2,组成向量 $(r_1, r_1 \times nu^1_\mu)$ 和 $(r_2, r_2 \times de^1_\mu)$. 计算 $r = r_2 / r_1$,把 r 发送到 S_0.

在分站点 S_0,有向量 $(nu^0_\mu, 1)$ 和 $(de^0_\mu, 1)$.

利用安全点积计算协议,计算 $nu_\mu = (r_1, r_1 \times nu^1_\mu) \cdot (mu^0_\mu, 1) = r_1 \times nu^0_\mu + r_1 \times nu^1_\mu = r_1 \times (nu^0_\mu + nu^1_\mu)$ 和 $de_\mu = (r_2, r_2 \times de^1_\mu) \cdot (de^0_\mu, 1) = r_2 \times de^0_\mu + r_2 \times de^1_\mu = r_2 \times (de^0_\mu + de^1_\mu)$.

在分站点 S_0,计算 $P_{\mu,\tau} = \bar{R}_\mu + r \times \frac{nu_\mu}{de_\mu}$.

4 安全性证明

4.1 安全两方计算[20]

我们考虑的可证安全性是建立在参与方是半诚实(semi-honest)前提下的. 半诚实参与方遵守协议规则,但它会根据协议执行过程中收到的结果试图破解,以获取额外的信息.

下面给出安全两方计算定义.

定义 1(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 的一个两方协议. 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 $(x, r, m_1, \cdots, m_t)$ 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 A_1 和 A_2,满足:

$\{A_1(x, f_1(x, y)), f_2(x, y)\}_{x,y \in \{0,1\}^*} \overset{c}{\equiv} \{VIEW_1^\Pi(x, y), OUTPUT_2^\Pi(x, y)\}_{x,y \in \{0,1\}^*}$

和 $\{A_2(y, f_2(x, y)), f_1(x, y)\}_{x,y \in \{0,1\}^*} \overset{c}{\equiv} \{VIEW_2^\Pi(x, y), OUTPUT_1^\Pi(x, y)\}_{x,y \in \{0,1\}^*}$,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.

基于定义 1,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(真实视图),也同样能够通过参与方的输入和输出在多项式时间内模拟获得. 也就是说,如果每一方都可以多项式时间地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.

可以看出,隐私保护两方协同过滤评分预测问题实质就是一个安全的两方计算问题:每一方拥有自己的本地隐私数据,都希望得到对某一指项的评分. 现在学界的很多数据分布存储的隐私保护数据挖掘研究或数学计算成果都是基于这个模型做出的. 下面给出文献[20]中已证明的半诚实模型下的合成定理.

定理 1(半诚实模型中的合成定理)

假设 g 可秘密约化为 f,且存在一个秘密计算 f 的协议,则存在一个秘密计算 g 的协议.

该定理实质是说,如果 g 可多项式时间内约化为 f,且存在一个隐私保护计算 f 的协议,则同样存在一个

隐私保护计算 g 的协议. 这是协议安全性证明的一个重要理论基础.

4.2 正确性证明

定理 2　在参与方是 2 的情况下,协议可以正确地计算出某指定用户(用评分向量表示)对指定项的评分.

证明:协议的正确性取决于第(4)步能否计算出对目标项的全局最近邻居相似度在两个分站点的本地最近邻居相似度下限值,以及第(5)步是否能计算到公式(2)中的 $\frac{\sum_{v\in NBS_u} sim(u,v)\times(R_{v,i}-\bar{R}_v)}{\sum_{v\in NBS_u} sim(u,v)}$.

在步(4)中,我们实际通过求两个分站点共 2×2^k 个数的最大的 2^k 个数来实现求 top-N 个数的目的[21,22]. 第一轮循环得到属于 top-2^k 的 2^k-1 个数,第二轮 2^k-2 个,…,最后一轮 1 个,加上循环外的 1 个,共 $2^k-1+2^k-2+\cdots+1+1=2^k$ 个. 这 2^k 个数就是两个分站点共 2×2^k 个最近邻居相似度中最大的 2^k 个. 这 2^k 个数包括了填入 S_1 中 B 数组的 (2^k-N) 个极大值,$2^k-(2^k-N)=N$,正好是 Top-N 最近邻居相似度. 所以第(4)步能计算对目标项的全局最近邻居相似度在两个分站点的本地分量下限值.

第(5)步利用了文献[10]中的两方安全相除协议,正确性比较明显:

$$P_{\mu,\tau}=\bar{R}_\mu+r\times\frac{nu_\mu}{de_\mu}$$

$$=\bar{R}_\mu+\frac{r_2}{r_1}\times\frac{r_1\times(nu^0_\mu+nu^1_\mu)}{r_2\times(de^0_\mu+de^1_\mu)}=\bar{R}_\mu+\frac{nu^0_\mu+nu^1_\mu}{de^0_\mu+de^1_\mu}=\bar{R}_\mu+$$

$$\frac{\sum_{v\in NBS^0_\mu} sim(u,v)\times(R_{v,\tau}-\bar{R}_v)+\sum_{v\in NBS^1_\mu} sim(u,v)\times(R_{v,\tau}-\bar{R}_v)}{\sum_{v\in NBS^0_\mu} sim(u,v)+\sum_{v\in NBS^1_\mu} sim(u,v)}$$

$$=\bar{R}_\mu+\frac{\sum_{v\in NBS_\mu} sim(u,v)\times(R_{v,\tau}-\bar{R}_v)}{\sum_{v\in NBS_\mu} sim(u,v)}$$

4.3 安全性证明

定理 3　在半诚实模型约束下,协议满足第 1 节中的安全性定义.

证明:在协议中,隐私数据交互发生在步(4)和步(5),因此证明这两步的安全性即可.

步(4):这实质上是个求 top-N 个数的协议,由合成定理,这个协议的安全性实际上取决于两方比较协议[23,24]的安全性. 注意到,安全两方比较协议应用到求 top-N 数值问题时,一方可以根据比较结果(0 或 1)猜测出另一方正在与之比较的数的值范围,要确保安全性,需证明任一方不能以大于 1/2 的概率推测出另一方正在比较的数的准确值. 下面是我们的证明:一个整数只有在比较两次的情况下,才有可能确定其值(否则只能确定值的范围). 一个数被比较两次的情况只能发生在数据列表只剩两个数的情况下. 不失一般性,假设分站点 S_0 正在比较的数据列表只剩两个数,是 $(a+2,a)$,而分站点 S_1 是 $(a+1,a-1)$. 首先 S_0 的 $a+2$ 和 S_1 的 $a+1$ 安全比较,输出为 0;接着 S_0 的 a 和 S_1 的 $a+1$ 比较,输出为 1. 分站点 S_0 猜测 S_1 中正在和它比较的数是 $a+1$,然而,如果 S_1 的数据列表是 $(a+2,a-1)$,也得到同样的输出结果,因此 S_0 不能判断 S_1 中正在和它比较的数是 $a+1$ 还是 $a+2$. 假设数出现的概率是相等的,S_0 不能以大于 1/2 的概率猜出数的准确值.

每分站点有 2^k 个数,设 S_0 的 2^k 个数组成集合 A,S_1 的 2^k 个数组成集合 B. 根据协议,要比较 $k+1$ 次. 第 i 次输出为 $\varepsilon_i\in\langle 0,1\rangle$,协议结束时,输出为 $(\varepsilon_0,\varepsilon_1,\cdots,\varepsilon_k)$,显然有:$f_0(A,B)=f_1(A,B)=output_1^{\mathrm{II}}(A,B)=output_1^{\mathrm{II}}(A,B)=(\varepsilon_0,\varepsilon_1,\cdots,\varepsilon_k)$.

那么 S_0 的视图 $View_0^{\mathrm{II}}(A,B)$ 是根据 ε_i 值对 S_1 中与之比较的 $k+1$ 个数的值的猜测(当最后两轮比较是 S_1 的同一个数时,是对 S_1 中 k 个数的值的猜测,S_0 可以根据自己参加比较的数的个数来判断. 如果 S_0 中有 k 个数,那么 S_1 中就有 $k+1$ 个;如果 S_1 中有 $k+1$ 个数,那么 S_0 中就有 k 个). S_1 的视图 $View_1^{\mathrm{II}}(A,B)$ 情况类似.

不影响证明的正确性和完整性,不妨假设 S_0 判断 S_1 中与它比较的是 k 个数,则:

$View_0^{\mathrm{II}}(A,B)=\{A,r^1,m_1^1,m_2^1,\cdots,m_k^1\}$. 其中 r_1 是 S_0 独立抛掷硬币的结果.

模拟器 A_0 根据 A 和输出 $(\varepsilon_0,\varepsilon_1,\cdots,\varepsilon_k)$ 得到的模拟视图为:

$A_0(A,f_0(A,B))=\{A,(r^1)',(m_1^1)',(m_2^1)',\cdots,(m_k^1)'\}$

注意到,$(m_1^1)',(m_2^1)',\cdots,(m_k^1)'$ 这 k 个数,和 S_0 的 $k+1$ 个数比较,产生输出 $(\varepsilon_0,\varepsilon_1,\cdots,\varepsilon_k)$. 由上面的分析知,站点 S_0 不能以大于 1/2 的概率猜对 S_1 中 k 个数的准确值,也就是说满足输出 $(\varepsilon_0,\varepsilon_1,\cdots,\varepsilon_k)$ 的 S_1 中参与比较的数所有可能值不止一组. 设:

$$\{m_1^1,m_2^1,\cdots,m_k^1\}=m\in M,\{(m_1^1)',(m_2^1)',\cdots,(m_k^1)'\}=m'\in M$$

M 是 S_1 中参与比较的数所有可能值的集合,由以上分析知 $|M|\geq 2$. 假设 m_r 为随机从 M 中抽取的元素,由于 m 和 m 也是 M 的随机观测元素,所以有:

$$Pr(m=m_r)=1/|M|=Pr(m'=m_r)$$

也就是说 m 和 m 是统计不可分的,根据文献[20]的统计不可分是计算不可分定理,可知 m 和 m 是计算不可分的. 进一步可得:

$\{A_0(A,f_0(A,B)),f_1(A,B)\}$

$\overset{c}{\equiv}\{ View_0^{II}(A,B), output_1^{II}(A,B)\}$

类似地可以构造模拟器 A_1,使得:

$\{ A_1(B, f_1(A,B)), f_0(A,B)\}$

$\overset{c}{\equiv}\{ View_1^{II}(A,B), output_0^{II}(A,B)\}$

由此得证第(4)步的安全性.

步(5):首先,分站点 S_0 收到从分站点 S_1 发过来的两个均匀分布随机数 r_1 和 r_2 相除的结果,设这个结果来自分布 D,那么很明显,从 D 中随机抽取一个数,就可以构成计算不可分的模拟视图了. 至于计算 nu_μ 和 de_μ 的安全性,依赖于两方点积计算的安全性,安全的两方点积计算协议很多,例如文献[9]和文献[10]中均有介绍. 在双方点积计算是安全的情况下,任一方只能得到点积结果,而不知道参与计算的向量元素值. 在本文的例子中,S_0 只能看到 nu_μ 和 de_μ 的值,而不知道 r_1、$r_1 \times nu_\mu^1$、r_2、$r_2 \times de_\mu^1$ 的值(r_1 和 r_2 在这里起到了"掩饰"S_1 中的数据的作用,但又不妨碍 S_0 得到最终结果). 不妨设 nu_μ 和 de_μ 的分布分别是 D_1 和 D_2,那么,随机从 D_1 和 D_2 中选取的两个数:nu_μ'、de_μ'、nu_μ、de_μ 分别是计算不可分的,可构成模拟视图,步(5)安全.

证明了步(4)和步(5)的安全性,由合成定理可知整个协议是安全的.

5 时间复杂度和通信耗费

协议的时间复杂度和通信耗费如表 1 所示. 其中:

t_1 是用户评分向量的编码位数;

N 是分站点 S_0 和 S_1 的用户数;

t_2 是邻居相似度值编码位数;

$l_1 = |NBS_\mu^0|$;

$l_2 = |NBS_\mu^1|$.

协议 5 的通信耗费,由安全两方点积计算决定,按照不同的准确度、安全性要求,有不同的通信耗费的安全两方点积计算协议,但一般都能达到多项式复杂度的通信耗费,设为 $O(P)$,时间复杂度为 $O(T)$.

表 1 协议各步时间复杂度和通信耗费

步骤	时间复杂度	通信耗费
1	常量	$O(t_1)$
2	$O(2 \times (N' + N' \times \log N'))$	无
3	$O(2 \times 2^k)$	无
4	$O(k)$	$O(2^{t_2} \times (k+1))$
5	$O(T + l_1 + l_2)$	$O(P)$

第(4)步的通信耗费是指数级的,这是由于采用文献[23]的安全两方比较协议的结果,如果使用文献[24]的安全两方比较协议,由于比较的通信耗费和位长成线性关系,为 $O(t_2)$,其通信耗费为 $O(t_2 \times (k+1))$.

6 结束语

本文针对参与方是 2 的情况,设计一个符合隐私定义的协同过滤预测评分安全协议,弥补了文献[1]中的协议不能应用于参与方是 2 的不足,使得分布式数据的隐私保护协同过滤推荐研究工作更加完整. 本文的安全两方协议,和文献[1]中的安全多方协议一样,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. 如何处理参与方不遵守半诚实约束的情况,文献[20]、[24]、[25]均提供了一些思路,是我们下一步的研究重点.

参考文献:

[1] 张锋,常会友. 基于分布式数据的隐私保持协同过滤推荐研究[J]. 计算机学报,2006,29(8):1487 - 1495.
Zhang Feng, Chang Hui-you. Research on privacy-preserving collaborative filtering recommendation based on distributed data [J]. Chinese Journal of Computers, 2006, 29(8): 1497 - 1495 (in Chinese).

[2] J Vaidya, C Clifton, M Zhu. Privacy Preserving Data Mining (Advances in Information Security) [M]. Springer-Verlag New York Inc, Nov. 2005.

[3] V S Verykios, E Bertino, et al. State-of-the-art in privacy preserving data mining [J]. SIGMOD Record, 2004, 33(1): 55 - 57.

[4] Y Lindell, B Pinkas. Privacy preserving data mining [A]. In: Advances in Cryptology-CRYPTO 2000, Proceedings of the 20th Annual International Cryptology Conference, LNCS 1880 [C]. 2000. 36 - 54.

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

[6] R Agrawal, R Srikant. Privacy-preserving data mining [A]. In: Proceedings of the 2000 ACM SIGMOD Conference on Management of Data [C], 2000. 439 - 450.

[7] 罗永龙,黄刘生等. 一个保护私有信息的布尔关联规则挖掘算法[J]. 电子学报,2005,33(5):900 - 903.
Luo Yong-long, Huang Liu-sheng, et al. An Algorithm for privacy-preserving boolean association rule mining [J]. Acta Electronica Sinica, 2005, 33(5): 900 - 903 (in Chinese).

[8] 黄毅群,卢正鼎,等. 分布式异常检测中隐私保持问题研究[J]. 电子学报,2006,34(5):796 - 799.
Huang Yi-qun, Lu Zheng-ding, et al. Privacy preserving outlier detection [J]. Acta Electronica Sinica, 2006, 34(5): 796 - 799 (in Chinese).

[9] C Clifton, M Kantarcioglu, et al. Tools for privacy preserving distributed data mining [J]. ACM SIGKDD Explorations, 2003,

4(2) :28 - 34.

[10] Du Wen-liang. A Study of Several Specific Secure Two-party Computation Problems [D]. PhD thesis, Purdue University, West Lafayette, Indiana, 2001.

[11] J Canny. Collaborative filtering with privacy[A]. In: Proceedings of the 2002 IEEE Symposium on Security and Privacy [C], Berkeley, 2002. 45 - 57.

[12] J Canny. Collaborative filtering with privacy via factor analysis [A]. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. 2002. 238 - 245.

[13] H Polat, Du Wen-liang. Privacy-preserving collaborative filtering using randomized perturbation techniques[A]. In: Proceedings of the Third IEEE International Conference on Data Mining[C]. 2003. 625 - 628.

[14] H Polat, Du Wen-liang. SVD-based Collaborative filtering with privacy[A]. In: Proceedings of the 20th ACM Symposium on Applied Computing, Track on E-commerce Technologies[C]. 2005. 791 - 795.

[15] Du Wen-liang., Zhan Zhi-jun. Using randomized response techniques for privacy-preserving data mining[A]. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. 2003. 505 - 510.

[16] L Sweeney. k-Anonymity: a model for protecting privacy[J]. International Journal on Uncertainty, Fuzziness and Knowledge-based System, 2002, 10(5) :557 - 570.

[17] L Sweeney. Achieving k-anonymity privacy protection using generalization and suppression, International Journal of Uncertainty[J]. Fuzziness, and Knowledge-based Systems. 2002, 10 (5) :571 - 588.

[18] C C Aggarwal. On the effects of dimensionality reduction on high dimensional similarity search[A]. In: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems[C]. 2001. 256 - 266.

[19] H Kargupta, S Datta, et al. Random data perturbation techniques and privacy preserving data mining[J]. Knowledge and Information Systems Journal, 2005, 7(4) :387 - 414.

[20] O Goldreich. Foundations of cryptography: volume 2, basic applications[M]. Beijing: Publishing House of Electronics Industry(Originated from Cambridge University Press), 2005.

[21] M Rodeh. Finding the median distributively[J]. Journal of Computer and Systems Sciences, 2004, 24:162 - 166

[22] G Aggarwal, Mishra N., Pinkas B.. Secure computation of the kth-ranked element[A]. In: Proceedings of IACR Advances in Cryptology(EUROCRYPT2004, LNCS 3027)[C]. 2004. 40 - 55

[23] A C Yao. How to generate and exchange secrets[A]. In Proc. 27th IEEE Sympsium on Foundations of Computer Science [C]. Washington, DC, USA: IEEE Computer Society Press, 1986. 162 - 167.

[24] C Cachin. Efficient private bidding and auctions with an oblivious third party[A]. In: Proceedings of the 6th ACM Conference on Computer and Communications Security[C]. 1999. 120 - 127.

[25] G Aggarwal, N Mishra, B Pinkas. Secure computation of the kth-ranked element[A]. In: Proceedings of IACR Advances in Cryptology-EUROCRYPT 2004, LNCS 3027[C]. 2004. 40 - 55.

作者简介:

张　锋　男,1974 年 7 月生于广西钦州. 毕业于中山大学计算机软件与理论专业,获博士学位. 现为中山大学软件学院讲师. 主要研究领域为隐私保护数据挖掘、机器学习算法、安全多方计算等.

E-mail: zhfeng@mail.sysu.edu.cn

孙雪冬　女,1972 年 10 月生于黑龙江阿城。毕业于哈尔滨工业大学计算机应用技术专业,获博士学位. 现为中山大学软件学院任讲师. 主要研究方向为工作流技术、复杂问题建模和优化等.

常会友　男,1962 年 11 月生于河北宣化. 毕业于哈尔滨工业大学计算机应用技术专业,获博士学位. 现为中山大学软件学院教授,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.

赵淦森　男,1977 年 9 月生于广东东莞,毕业于英国肯特大学(University of Kent, UK)计算机软件与理论专业,获博士学位. 现为中山大学软件学院教师. 主要研究领域为安全协议分析、信任管理、密码系统、网络安全等.