

一种带隐私保护的基于标签的推荐算法研究

曹春萍 徐帮兵

(上海理工大学光电信息与计算机工程学院 上海 200093)

摘要 在基于标签的推荐中,标签起着联系用户和信息资源的作用。但由于存在语义特性,相较于评分数据,标签数据在一定程度上更能够直接反映用户喜好,隐私问题更为突出。推荐服务器收集用户的历史标签记录,一旦攻击者通过攻击推荐服务器而获得了用户信息,将造成严重的用户隐私泄露问题。对此,提出一种带有隐私保护的基于标签 k-means 聚类的资源推荐方法 CDP k-meansRA,即利用 Crowds 网络进行用户发送方匿名保护,并且将 ϵ -差分隐私保护融入改进的标签 k-means 聚类算法中。通过实验将提出的 CDP k-meansRA 与 k-meansRA 等算法进行比较,证明了 CDP k-meansRA 能够在保护用户隐私的前提下,保证一定的推荐质量。

关键词 隐私保护,标签聚类,Crowds 网络,发送方匿名, ϵ -差分隐私

中图分类号 TP309.2 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.08.024

Research of Privacy-preserving Tag-based Recommendation Algorithm

CAO Chun-ping XU Bang-bing

(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract In tag-based recommendation, tags play a role in the link between users and information resources. However, compared to rating data, since the semantic properties of the tag data, tag data reflects user preferences more directly, so the privacy issues in tag-based recommendation are more serious. Recommender server collects user history tag records, once an attacker accesses the user information by attacking the recommender server, it will cause serious leakage of user privacy. A resource recommendation method (CDP k-meansRA) based on tag k-means clustering with privacy protection is proposed. Sender anonymity protection is provided by using Crowds network and ϵ -differential privacy are fused into an improved tag clustering based recommendation algorithm. The experiments show that compared to k-meansRA and so on, the CDP k-meansRA can keep the quality of the recommendation under the premise of user privacy preservation.

Keywords Privacy preservation, Tag clustering, Crowds network, Sender anonymity, ϵ -differential privacy

1 引言

互联网的快速发展使得信息过载成为一个亟需解决的问题,个性化推荐服务成为解决信息过载的一种重要手段。个性化推荐服务大大方便了人们的生活,但由于个性化推荐系统需要收集用户数据信息并分析挖掘用户兴趣模型,而用户数据信息包含了大量个人隐私,因此如何防止用户隐私泄露成为了推荐领域的一大研究热点。

目前,推荐系统的隐私保护方法一般分为数据加密、数据泛化和数据扰动 3 类^[15]。文献[1-2]采取 k-anonymity 模型将用户标识信息泛化处理,要求对于任一条记录,与其具有相同邻居敏感标签的信息至少有 $k-1$ 个,将用户信息隐匿于 k 个同类中。但 k 匿名隐私保护方法并没有对攻击者掌握的背景知识进行定义,因此总是需要因新型攻击的出现而不断完善^[9]。同态加密技术在安全多方计算协同过滤中得到使用,

文献[3]提出了一种基于加密的隐私保护协同过滤中的安全两方协议;文献[4]提出了一种分量形式为二次方的同态加密公钥生成方法,但加密算法本身十分复杂,生成的公钥尺寸太大,并且没有对隐私做出严格定义;文献[7,11]提出了扰动强度的概念以及度量方法,对基于扰动技术的推荐方法进行改进,虽然数据扰动方法较为简便,但其存在保护能力不强的问题^[6]。

以上研究方法并没有对攻击者的背景知识进行规范约束,往往随着攻击者掌握背景知识的增强会出现新的攻击方式,此时需要研究新的对应保护模型。相较于传统隐私保护方法, ϵ -差分隐私定义了一个严格的、可证明的隐私保护模型^[9]。目前 ϵ -差分隐私应用于推荐系统方面的研究和成果较少,文献[16]最早将 ϵ -差分隐私引入项目协方差矩阵中,随后在矩阵上利用传统无隐私保护推荐算法进行评分预测;文献[17]进一步将 ϵ -差分隐私运用到评分矩阵分解模型中,以达

到稿日期:2016-07-31 返修日期:2016-12-03 本文受国家自然科学基金项目(61202376),上海市自然科学基金(15ZR1429100)资助。

曹春萍(1968-),女,硕士,副教授,主要研究方向为智能决策支持系统、个性化服务;徐帮兵(1991-),男,硕士生,主要研究方向为数据挖掘、推荐系统,E-mail:bangbingb@126.com。

到隐私保护的效果。但是这些研究大多针对评分预测情况,而对于标签数据隐私保护的研究几乎处于空白。相较于评分数据隐私,基于标签的推荐隐私问题更为突出,因为标签语义的特性使其在一定程度上能够直接反映出用户的兴趣,一旦被攻击者获得将造成用户隐私泄露问题。

对此,本文提出了一种能够提供用户发送方匿名保护并且满足 ϵ -差分隐私保护的推荐算法,以达到在不严重影响推荐性能的前提下,提供用户隐私保护的目标。在服务器数据收集阶段,利用 Crowds 网络对用户进行匿名发送保护,使得推荐服务器无法识别用户身份;在资源推荐预测阶段,将 ϵ -差分隐私引入标签聚类过程中,使得攻击者无法推测出用户的数据信息,将用户概貌表示成基于聚类空间的向量形式,并利用内容推荐算法完成推荐。

2 相关理论

相较于传统隐私保护方法存在的安全性与攻击者掌握的背景知识相关以及无法严格证明其隐私保护水平这两个缺陷, ϵ -差分隐私是一个与背景知识无关,并且严格可证的隐私保护模型^[9]。它的严格可证是建立在其严格的定义和坚实的数学基础上的。 ϵ -差分隐私保护模型假设攻击者可以掌握最大背景知识,例如攻击者已获得除目标记录以外的其他所有记录信息,因此 ϵ -差分隐私无需再考虑攻击者背景知识的掌握程度;并且, ϵ -差分隐私有着严格的数学定义,对隐私保护进行了量化评估,使得隐私保护水平能够通过隐私预算参数比较。 ϵ -差分隐私模型由 Dwork 团队提出^[5,8,13],本文先简单介绍 ϵ -差分隐私的相关理论,后文将会依赖这些理论对算法进行阐述。下面给出 ϵ -差分隐私的一些核心概念及定义。数据集 D 和 D' 具有相同的属性结构,若对称差大小 $|D \Delta D'|=1$,则 D 和 D' 是邻近数据集。

定义 1(差分隐私, Differential Privacy) 设数据集 D 和 D' 是邻近数据集,设定的随机算法为 M , P_M 为 M 的所有输出构成的集合, M 在 D 和 D' 的任意输出结果为 S_M , 若 M 满足:

$$P_r[M(D) \in S_M] \leq \exp(\epsilon) \times P_r[M(D') \in S_M] \quad (1)$$

则算法 M 满足 ϵ -差分隐私保护。其中 ϵ 称为隐私保护预算,表示隐私保护程度。 ϵ 值越大,保护程度越低; ϵ 值越小,保护程度越高;若 $\epsilon=0$,则表示保护程度最大。

实现 ϵ -差分隐私保护的主要技术是噪声的添加。如何添加噪声使算法满足 ϵ -差分隐私保护,与函数的敏感度以及 ϵ 隐私预算直接关联。

定义 2(函数的敏感度) 对于函数 $f: D \rightarrow R^d$, 其输入为数据集 D , 输出为 d 维实数, 则函数 f 的敏感度 Δf 为:

$$\Delta f = \max_{D, D'} \|f(D) - f(D')\| \quad (2)$$

其中, D' 为 D 的邻接数据集。

在实践中,为使一个算法满足 ϵ -差分隐私保护,常用到拉普拉斯机制。拉普拉斯机制是通过添加服从拉普拉斯分布的随机噪声来实现 ϵ -差分隐私保护的。拉普拉斯分布的概率密度函数为:

$$f(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right) \quad (3)$$

其中, μ 为位置参数, b 为尺度参数。

定义 3(拉普拉斯机制) 对于数据集 D , 给定函数 $f: D \rightarrow R^d$, 记函数敏感度为 Δf , 则随机算法:

$$M(D) = f(D) + X \quad (4)$$

提供了 ϵ -差分隐私保护, 其中 $X \sim \text{Lap}\left(\frac{\Delta f}{\epsilon}\right)$ 为随机噪声, 服从尺度参数为 $\frac{\Delta f}{\epsilon}$ 的拉普拉斯分布。

ϵ -差分隐私模型具有以下两条特性。

性质 1(序列组合特性) 假设随机算法组合 $M = \{M_1, M_2, \dots, M_n\}$, 对于同一数据集 D , M_i 提供 ϵ_i -差分隐私, M 提供 $\sum_{i=1}^n \epsilon_i$ -差分隐私。

性质 2(并行组合特性) 假设随机算法组合 $M = \{M_1, M_2, \dots, M_n\}$, 对于数据集 D 的不相交的子集各自满足 ϵ_i -差分隐私保护, 则 M 对 D 提供 $\max \epsilon_i$ -差分隐私保护。

3 算法介绍

推荐服务器为提供推荐服务收集用户行为信息, 由于数据完全存储在服务器上, 一旦服务器数据泄露将造成严重的隐私泄露问题; 而在推荐阶段, 攻击者结合一定的背景知识将能够推断出用户行为数据。因此, 本文提出的算法分为两步: 1) 进行用户标签信息匿名提交, 利用 Crowds 通信网络提供发送方匿名保护, 由于数据接收方即推荐服务器是不需要匿名的, 因此只需要满足发送方匿名即可; 2) 进行标签聚类, 将用户概貌和资源映射成标签类簇表示的方式, 并且将 ϵ -差分隐私保护引入聚类过程, 从而防止攻击者推测出用户的原始数据。

3.1 用户信息匿名提交

推荐系统需要收集用户个人信息, 服务器可获得完全的个人信息, 而且服务器管理员将能够轻易获得用户数据, 这些数据一旦泄露, 将造成严重的隐私泄露问题^[15]。因此, 本文提出用户信息的匿名提交方法, 以对用户进行匿名保护。

Crowds 网络基于重路由机制提供信息提交匿名保护, 一条重路由路径可以通过以下方式表示^[12]:

$$\langle S, N_1, N_2, \dots, N_n, R \rangle \quad (5)$$

其中, S 表示信息发送方, R 表示信息接收方, N_i 表示中继节点, n 表示中继节点个数。信息发送方需要发送信息至接收方时, 会随机建立一条发送路径, 过程如下: 发送方选择一个中继节点作为代理节点并将数据发送给中继节点, 中继节点接收数据后, 可以选择发送给下一个中继节点或者直接发送给信息接收方。接收方接收信息后如需应答, 则通过同一路径返回应答信息。路径中的每个节点都只知道自己的前驱节点和接收方, 即发送方匿名。Crowds 网络的发送方匿名特性十分契合用户标签数据匿名提交的需要。针对标签系统, 用户发送方信息匿名提交的过程如下:

(1) 用户节点发送数据 (address, data) 给中继代理节点, 其中 address 为当前发送方地址, 在此即为用户地址; data 为标签数据集, 包含各个 (tag, resource) 标签-资源数据对。

(2) 中继代理节点记录 address, 并修改 address 地址为自己的地址, 同时以预设值概率 p 选择是否传输给下一个中继节点, 若继续传输给下一个中继则重复步骤 (2), 否则将数据信息发送给服务器。

(3)服务器接收信息后,得到节点地址 address,同时保存 data 信息以便后续处理,若需返回回答信息,则发送给 address 节点。每个节点根据记录的 address 地址依次返回回答信息,从而将信息原路径返回。

标签数据提交后,推荐服务器并不知道数据的来源,从而达到用户发送方匿名保护。推荐服务器通过 Crowds 收集数据后,将进行隐私标签聚类。

3.2 隐私保护下的标签聚类

内容推荐算法将用户兴趣构建成一个空间向量模型,同时将资源特征表示成空间向量模型,并进行相似度计算,产生资源推荐列表。在基于标签推荐中,利用标签作为用户兴趣以及资源特征的表示中介,可以将用户兴趣表示为如下形式:

$$anonym_u = \{T_n, W_n\} \quad (6)$$

其中, $anonym_u$ 表示某一匿名用户, T_n 和 W_n 分别表示标签以及对应的权重, n 为系统内标签的数量。资源的表示同样采用上述方式。

差分隐私是基于数据失真技术来进行隐私保护的^[9]。如果向用户兴趣空间向量模型直接添加随机噪声,将产生大量的噪声,虽然隐私得到了保护,但是推荐系统的性能将严重下降。为此,我们采用了结合差分隐私的标签聚类算法,并以此提出一种带有隐私保护的推荐算法 CDP k-meansRA。该算法将匿名用户兴趣以及资源的表示转化成基于类簇的表示形式,以达到降噪效果,并且满足差分隐私保护。如图 1 所示,攻击者即使从其他渠道获得了最大背景知识,拥有和数据集 D 仅相差一条记录的近邻数据集 D' ,根据计算结果也无法推断出这条记录的隐私信息。

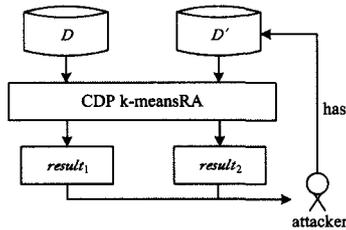


图 1 应对攻击模型

文献[10]提出了一种区别于传统的基于标签共现(tag co-occurrence)聚类算法——基于对象特征向量表示法的标签聚类算法。文中提出 3 种不同的标签表征方法:基于资源的特征向量(Item-Based-Vector, IBV)、基于其他共现标签的特征向量(Tag(Others)-Based-Vector, TOBV)和基于全集共现标签的特征向量(Tag(All)-Based-Vector, TABV),并通过实验得出 IBV 和 TABV 的性能更优。本文拟采用 TABV 方法来表示标签,然后将差分隐私融入 k-means 聚类算法,得到标签类簇。

3.2.1 标签的特征向量表示及相似度计算

为了将标签进行 k-means 聚类运算,首先需要定义标签的特征向量表示,本文采用 TABV 来进行标签的特征向量表示。标签 $t_i \in T$ 的特征向量如式(7)所示:

$$V_{t_i} = (w_{t_i,1}, w_{t_i,2}, \dots, w_{t_i,|T|}) \quad (7)$$

其中, $w_{t_i,j}$ 如式(8)所示:

$$w_{t_i,j} = \begin{cases} |\{k|b_{k,i} > 0\}|, & \text{if } i=j \\ |\{k|b_{k,i} > 0 \& b_{k,j} > 0\}|, & \text{if } i \neq j \end{cases} \quad (8)$$

其中, $|T|$ 表示标签的总数, $b_{k,i}$ 表示资源 r_k 被标签 t_i 标注的次数。TABV 的思想是将标签特征向量元素值表示为与其他标签的共现次数,并且将自身元素项的值设为该标签标注的资源数。在实际情况中使用该方法表示标签时,特征向量具有稀疏性的特点,从而保证了聚类算法的时间和空间复杂度不会过高。

本文采用余弦公式计算标签距离:

$$d(t_i, t_j) = 1 - sim(t_i, t_j) = 1 - \frac{V_{t_i} \cdot V_{t_j}}{|V_{t_i}| \times |V_{t_j}|} \quad (9)$$

3.2.2 隐私保护下的标签聚类

为了减少添加的随机噪声量,同时为了减轻标签模糊问题造成的推荐效果不佳的问题,进行标签聚类操作。本文提出的算法是基于 k-means 聚类算法的,同时将 ϵ -差分隐私引入聚类过程中,以达到隐私保护效果。k-means 算法在计算样本点与中心点的距离时会泄露隐私^[14]。因此,可以通过隐藏类簇的中心和标签数量来防止攻击者推断出标签所属的类簇分组。k-means 聚类操作的一般过程如下。

算法 1 k-means 算法

1. 给定数据集 D , 包含 n 个 d 维空间向量样本点;
2. 随机选取 k 个随机样本点作为初始的类簇中心;
3. 计算其他样本点到中心点的距离,从而选择最近距离归类;
4. 根据公式 $c_i = \frac{\sum_{p_j \in C_i} p_j}{num_i}$ 重新计算样本中心点,其中 num_i 为类簇 C_i 中的样本数量, p_j 为类簇 C_i 中的样本点;
5. 如果中心点不变或者达到迭代次数,则算法结束;否则,跳转到第 3 步执行。

为了达到隐私保护效果,需要隐藏标签类簇的中心以及标签数量,本文采用拉普拉斯随机机制添加噪声,在计算样本中心点时,对分子求和并将分母计数结果加入拉普拉斯噪声。隐私总预算为 ϵ , 函数敏感度 Δf 与函数相关,计数 num_i 的敏感度显然为 $\Delta f = 1$ 。对于 d 维空间向量,属性求和的最小值为 0,最大值为 1,故 d 维求和敏感度为 $\Delta f = d$ 。因此两者序列组合敏感度 $\Delta f = d + 1$,如果迭代次数固定为 p ,添加噪音为 Laplace($\frac{p(d+1)}{\epsilon}$),则类簇中心点的计算公式为:

$$c_i = \frac{\sum_{p_j \in C_i} p_j + \text{Laplace}(\frac{2p(d+1)}{\epsilon})}{num_i + \text{Laplace}(\frac{2p(d+1)}{\epsilon})} \quad (10)$$

如果迭代次数未给定,则可以采取文献[14]中给出的预算减半策略,即每一次迭代都采用剩余预算的一半,则第 i 次迭代时的隐私预算为 $\epsilon_i = (\epsilon/2)^i$,其中 ϵ 为算法总预算。

由于初始标签中心点的选择对最终的聚类效果影响较大,因此不采用随机选择的方法。考虑到聚类应该达到类内部相似度尽量高而类间相似度尽量低的效果,在选取初始标签类簇中心时应尽量选取彼此间相似度较小的标签。假设有一标签集合 $Z = \{t_1, t_2, \dots, t_{|Z|}\}$,给出集合外的一个标签 t_{out} 到标签集合 Z 的距离计算公式:

$$d(t_{out}, Z) = \sum_{t_i \in Z} d(t_{out}, t_i) \quad (11)$$

其中, $d(t_{out}, t_i)$ 采用式(9)进行计算。结合 ϵ -差分隐私,在计算 $d(t_{out}, Z)$ 时加入拉普拉斯噪声。根据定义 2,敏感度与具体函数相关, $d(t_{out}, Z)$ 的敏感度 $\Delta f = \max d(t_{out}, t_i)$,而

$d(t_{out}, t_i)$ 采用余弦公式计算,因此最大差值为 1。所以函数敏感度为 1,隐私分配预算为 $\frac{\epsilon}{p \cdot (k-1)}$,因此 ϵ -差分隐私下标签 t_{out} 到标签集合 Z 的距离计算公式为:

$$d'(t_{out}, Z) = \sum_{t_i \in Z} d(t_{out}, t_i) + \text{Laplace}\left(\frac{p \cdot (k-1)}{\epsilon}\right) \quad (12)$$

其中, p 为迭代次数, d 为标签向量维数, k 为初始类簇中心标签数量。初始类簇中心标签的选择依据算法 2 进行。

算法 2 初始标签中心选择

1. 给定 TABV 表示的标签集合 $T = \{t_i | 1 \leq i \leq |T|\}$, t_i 为 TABV 表示的标签;
2. 利用公式 $c_1 = \frac{\sum t_i}{|T|}$ 计算得到第一个类簇中心标签 c_1 , 并将其加入到中心标签集合中, $C = \{c_1\}$;
3. 利用式(12)计算中心集合 C 外的所有标签到集合 C 的距离 $d'(t_{out}, C)$, 选取 d' 最大的标签 t_i 加入中心标签集合中, $C = \{c_1, t_i\}$;
4. 如果 $|C| < K$, 则重复第 3 步直至选取出 k 个初始类簇中心标签。

本文的差分隐私标签聚类的步骤如算法 3 所示。

算法 3 差分隐私标签聚类

1. 给定 TABV 表示的标签集合 $T = \{t_i | 1 \leq i \leq |T|\}$, t_i 为 TABV 表示的标签;
2. 利用算法 2 得到 k 个初始类簇中心标签;
3. 计算其他标签到中心标签的距离, 从而选择最近距离的中心标签归类;
4. 如果标签聚类类簇结果不变或者达到预定迭代次数, 则结束, 否则利用式(10)计算 k 个类簇中心标签, 并跳转到第 3 步执行。

3.3 完成个性化资源推荐

在进行标签聚类后, 一定程度上将语义相似的标签放在了一个类簇之中, 不同类簇可以视为不同的主题。我们计算出匿名用户对于各个类簇的兴趣权重, 从而将匿名用户的概貌用基于类簇的向量模型进行表示:

$$\text{anonym}_u = \{TC_k, W_k\} \\ = \{(tc_1, w_{1,u}), (tc_2, w_{2,u}), \dots, (tc_k, w_{k,u})\} \quad (13)$$

其中, tc_i 表示第 i 个类簇, w_i 表示对应类簇的兴趣权重, 采用如下方式计算:

$$\text{anonym}_u(w_i) = \frac{|t_u \cap tc_i|}{|t_u|} \quad (14)$$

其中, t_u 表示用户 u 使用的标签集合。同样, 将资源也表示成基于类簇的模型向量:

$$r = \{TC_k, W_{r,k}\} \\ = \{(tc_1, w_{r,1}), (tc_2, w_{r,2}), \dots, (tc_k, w_{r,k})\} \quad (15)$$

其中, $w_{r,i}$ 表示资源 r 在类 tc_i 上的相关性权重, 采用如下公式计算:

$$w_{r,i} = \frac{|t_r \cap tc_i|}{|t_r|} \quad (16)$$

其中, t_r 表示标注在资源 r 上的标签集合。匿名用户 u 对资源 r 的兴趣值采用式(9)计算, 将其降序排列获得 TOP-N 列表, 按照请求原路径将响应结果返回给匿名用户, 并剔除已有过行为的资源。

3.4 关于算法满足 ϵ -差分隐私的证明

上述算法主要通过添加拉普拉斯噪声来实现 ϵ -差分隐私保护, 因此需要证明该算法是严格遵守 ϵ -差分隐私定义的。下面将对算法的每一个步骤进行分析, 并根据第 2 节的定义

以及性质进行证明。

1) 首先利用 Crowds 网络进行用户发送方匿名保护, 这一步骤没有采用 ϵ -差分隐私。

2) 算法 2 进行初始类簇中心标签的选取, 在计算标签到中心标签集合的距离时, 除了第一个中心标签之外, 其他每个中心标签的计算都加入了拉普拉斯噪声, 根据定义 3 可知这 $k-1$ 个标签都满足 $\frac{\epsilon}{p \cdot (k-1)}$ -差分隐私, 并且在算法 2 中计算中心标签时是串行逐一计算的, 具有序列组合性, 由性质 1 可知算法 2 的初始中心标签的选择满足 $\sum_{i=1}^{k-1} \frac{\epsilon}{p \cdot (k-1)} = \frac{\epsilon}{p}$ -差分隐私。

3) 算法 3 中, 在同一次迭代计算中心标签过程内, 每一个中心标签都满足 $\frac{\epsilon}{p}$ -差分隐私, 并且由于每个中心标签的计算都在各自独立的类簇中进行的, 因此一次迭代计算是满足并行组合性的, 由性质 2 可知 k 个中心标签依然满足 $\frac{\epsilon}{p}$ -差分隐私。由于一次迭代计算 k 个中心标签是在上一次迭代计算之后进行的, 即多次迭代计算属于串行运算, 因此具有序列组合性。由性质 1 知其满足 $\frac{(p-1)\epsilon}{p}$ -差分隐私。

4) 算法 2 步骤是算法 3 迭代计算的前置处理, 即两者是串行计算的, 因此满足序列组合性。由性质 1 可知, 算法整体满足 $\frac{\epsilon}{p} + \frac{(p-1)\epsilon}{p} = \epsilon$ -差分隐私。

因此本文算法满足 ϵ -差分隐私。

4 实验与结果分析

本节将在真实数据集上对本文提出的算法进行实验, 并对实验结果进行观察分析, 以验证所提算法的效果。

4.1 实验数据集

本文选择在 CiteULike 标签数据集¹⁾上进行实验。CiteULike 系统是典型的社会标签系统, 用户能够收藏论文, 并对其标注标签。CiteULike 数据集是其公布的一个真实数据集, 能为许多领域的研究者提供重要的研究价值。为了避免数据集中数据稀疏问题对实验的影响, 这里只考虑数据集中相对稠密的部分数据集。经过整理, 实验部分数据集包含 1326 位用户, 使用标签共 14396 条, 资源数量共 4011 个, 标签标注行为 100566 次。

4.2 聚类 k 值

衡量聚类算法性能的一个流行方法是计算轮廓系数(Silhouette Coefficient)。一个聚类算法的目标是使聚类结果的类间分离度较高, 并且类内部的内聚度较大, 而轮廓系数同时考虑了内聚度和分离度。因此, 本文将采用轮廓系数进行 ϵ -差分隐私标签聚类算法的性能分析。标签 t_i 的轮廓系数 $SC(t_i)$ 的计算公式如下:

$$SC(t_i) = \frac{b(t_i) - a(t_i)}{\max(b(t_i), a(t_i))} \quad (17)$$

其中, $a(t_i)$ 表示标签 t_i 到其所属标签类簇中其他标签的平均距离, $b(t_i)$ 表示标签 t_i 到其他各个标签类簇中标签平均距离

¹⁾ <http://www.citeulike.org>

的最小值。轮廓系数值的范围是 $[-1, 1]$ 。

ϵ -差分隐私标签聚类的性能将使用所有标签的平均轮廓系数来衡量:

$$S = \frac{1}{|T|} \sum_{t_i \in T} \frac{b(t_i) - a(t_i)}{\max(b(t_i), a(t_i))} \quad (18)$$

轮廓系数越大,则聚类性能越好。

首先采用上文所述的 TABV 对标签进行表示,并进行归一化处理。将对原始 k-means 算法、通过计算选取初始类中心标签的 C k-means 算法、随机选取初始类簇中心的 ϵ -差分隐私聚类算法 RDP k-means 以及利用算法 3 进行的通过计算来选取类簇中心的 ϵ -差分隐私标签聚类算法 CDP k-means (calculated-centers differential privacy k-means) 这 4 种算法进行对比。固定隐私预算 $\epsilon=1$,通过实验来观察 k 值对聚类结果的影响。在实验中将 k 值设为 $[5, 60]$ 内的整数,实验结果如图 2 所示。

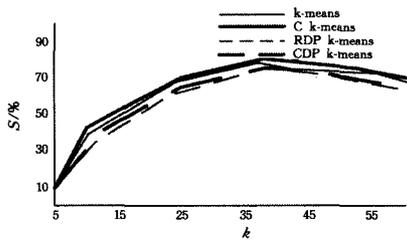


图2 轮廓系数与 k 值

如图 2 所示,随着 k 值的增加,4 种算法的平均轮廓系数 S 在总体上呈上升趋势,当 k 在 $[35, 40]$ 之间时, S 值较大,此时聚类效果较好。CDP k-means 算法聚类效果在总体上优于 RDP k-means,并且 CDP k-means 的聚类性能在 $\epsilon=1$ 的情况下十分接近非隐私保护下的 C k-means 的性能,比如在 $k=39$ 时, $S(C k-means) = 0.0710$, $S(RDP k-means) = 0.0690$, $S(CDP k-means) = 0.0703$,对于 C k-means 来说,CDP k-means 性能只下降了 0.9%;而相对于 RDP k-means 来说,CDP k-means 性能上升了 1%左右。综合考虑,取 k 值为 36,此时 CDP k-means 较 C k-means 性能下降不到 1%,同时也接近最佳聚类的 k 值。关于 ϵ 值对性能的影响,我们将其作为推荐性能的观察变量来进行实验。

4.3 隐私预算 ϵ

本文将使用 F 值来验证提出的基于 ϵ -差分隐私标签聚类的推荐算法的性能。F 值可以综合考虑准确率及召回率。准确率 precision 表示系统给出的推荐列表中预测正确的资源数量占据资源预测数量的比例,而召回率 recall 表示系统给出的推荐列表中预测正确的资源数量占据测试数据集中用户使用资源数量的比例,两者分别用以下公式来定义:

$$precision = \frac{1}{|U|} \sum_{u \in U} \frac{|hits_u|}{|rec_u|} \quad (19)$$

$$recall = \frac{1}{|U|} \sum_{u \in U} \frac{|hits_u|}{|test_u|} \quad (20)$$

其中, $hits_u$ 为命中的资源集合, $test_u$ 为测试集中用户 u 使用的资源集合, rec_u 表示推荐资源集合, U 为用户集合。

F 值为准确率和召回率的调和平均值,计算公式为:

$$F = \frac{2 \times precision \times recall}{precision + recall} \quad (21)$$

F 值越大,说明推荐性能越好。

本文将数据集分为两部分,80%的数据集作为训练集,20%的数据集作为测试集。我们观察隐私预算 ϵ 值在 $0.1 \sim 1$ 内的推荐效果,推荐列表长度设为 50 时,k-meansRA, C k-meansRA, RDP k-meansRA 以及 CDP k-meansRA 的实验结果如图 3 所示。

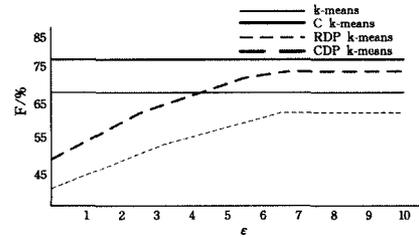


图3 F 值与隐私预算 ϵ 值的关系

如图 3 所示,随着隐私预算 ϵ 的增加,推荐算法 RDP k-meansRA 和 CDP k-meansRA 的 F 值都呈上升趋势,并且 $\epsilon < 0.65$ 时,两个算法的 F 值相对较低 $\epsilon > 0.65$ 时,F 值趋于稳定。这是由于当 ϵ 预算变大之后,隐私保护力度变小,所添加的噪声变小,因此 F 值升高,并在 $\epsilon > 0.65$ 之后,F 值上升趋势趋于稳定。将 CDP k-meansRA 与 C k-meansRA 进行比较可知,在 $\epsilon \in (0, 0.65)$ 时,尤其是在 $\epsilon < 0.3$ 时,CDP k-meansRA, RDP k-meansRA 隐私保护推荐算法的 F 值严重低于 C k-meansRA 和 k-meansRA,这表明如果隐私保护力度过大,即 ϵ 过小,将会使得推荐性能严重下降。当 $\epsilon = 0.7$ 时, $F(CDP k-meansRA) = 74.3\%$,而 $F(C k-meansRA) = 76\%$, $F(k-meansRA) = 66\%$,此时 CDP k-meansRA 较 C k-meansRA, F 值只下降了 2.2%,处于可接受范围;CDP k-meansRA 较 k-meansRA, F 值提升了 12.5%。实验结果证明,本文提出的隐私保护下的基于标签聚类的推荐算法不仅能够一定程度上保护用户数据隐私,并且在推荐性能上没有大幅度下降。

结束语 用户在资源上标注标签,无疑使得标签成为良好的联系用户和资源的桥梁,但是服务器在用户数据收集以及产生推荐过程中的用户数据隐私问题十分突出。为实现用户数据匿名提交,本文利用 Crowds 网络的发送方匿名特性对数据匿名提交提供支持。k-anonymity 以及同态加密等隐私保护技术都没有对隐私作出严格定义,并且保护力度与攻击者掌握的背景知识相关,而 ϵ -差分隐私模型克服了这些缺点。因此本文将 ϵ -差分隐私保护模型融入标签聚类推荐算法中,对推荐过程提供隐私保护,并在真实数据集上通过实验验证了算法的性能。

参考文献

- [1] BILENKO M, RICHARDSON M. Predictive client-side profiles for personalized advertising[C]//Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining(KDD'11). San Diego, California, USA, 2011:413-421.
- [2] CHEN C L, XIONG J, CHEN L, et al. Personalized Privacy Preservation Algorithm in Weighted Social Networks[J]. Computer Technology and Development, 2016, 26(8): 88-92. (in Chinese) 陈春玲,熊晶,陈琳,等.加权社会网络中的个性化隐私保护算法[J].计算机技术与发展,2016,26(8):88-92.
- [3] LIU H W, LIANG F, ZHU H. On secure two-party protocol for privacy protection recommendation system[J]. Computer Appli-

- cations and Software, 2014, 31(8): 17-19. (in Chinese)
- 刘洪伟, 梁飞, 朱慧. 面向隐私保护推荐系统的安全两方协议研究[J]. 计算机应用与软件, 2014, 31(8): 17-19.
- [4] CRONJ S, MANDALA N. Fully homomorphic encryption over the integers with shorter public keys [M]// *Advances in Cryptology-CRYPTO 2011*. Berlin: Springer, 2011: 94-145.
- [5] DWORK C. Difference Privacy[C]// *Proc of the 33rd International Colloquium on Automata, Languages and Programming, Part II*. 2006: 1-12.
- [6] PENG F, ZENG X W, LIU L, et al. Privacy preserving recommendation method based on groups[J]. *Application Research of Computers*, 2015, 32(3): 869-871. (in Chinese)
- 彭飞, 曾学文, 刘磊, 等. 一种基于群组推荐的用户隐私保护方法[J]. 计算机应用与研究, 2015, 32(3): 869-871.
- [7] ZHANG F Z, LIU T, FENG S S. Improved Privacy-preserving Collaborative Filtering Recommendation Algorithm [J]. *Computer Engineering*, 2010, 36(16): 126-134. (in Chinese)
- 张付志, 刘亭, 封素石. 一种改进的隐私保持协同过滤推荐算法[J]. 计算机工程, 2010, 36(16): 126-134.
- [8] DWORK C. Difference privacy: a survey of results[C]// *Theory and Applications of Models of Computation*. 2008: 1-9.
- [9] XIONG P, ZHU T Q, WANG X F. A Survey on Differential Privacy and Applications[J]. *Chinese Journal of Computers*, 2014, 37(1): 101-122. (in Chinese)
- 熊平, 朱天清, 王晓峰. 差分隐私保护及其应用[J]. 计算机学报, 2014, 37(1): 101-122.
- [10] ZHOU J, CHEN C, YU N H. Tag Clustering Algorithm Using Object-based Feature Vector[J]. *Journal of Chinese Computer Systems*, 2012, 33(3): 525-530. (in Chinese)
- 周津, 陈超, 俞能海. 采用对象特征向量表示法的标签聚类算法[J]. 小型微型计算机系统, 2012, 33(3): 525-530
- [11] AGGARWAL C C. On randomization, public information and the curse of dimensionality[C]// *Proc of the 23rd IEEE International Conference on Data Engineering*. IEEE Computer Society, 2007: 136-145.
- [12] ZHANG J C, TANG X, SUN Y. Study of Crowds Anonymous Communication System[J]. *Microcomputer Information*, 2012, 28(7): 128-130. (in Chinese)
- 章敬崇, 唐旭, 孙宇. Crowds 匿名通信系统研究[J]. 微计算机信息, 2012, 28(7): 128-130.
- [13] DWORK C. A Firm Foundation for Private Data Analysis[C]// *Communications of the ACM*. 2011: 86-95.
- [14] LI Y, HAO Z F, WEN W, et al. Research on Differential Privacy Preserving k-means Clustering[J]. *Computer Science*, 2013, 40(3): 287-290. (in Chinese)
- 李杨, 郝志峰, 温雯, 等. 差分隐私保护 k-means 聚类方法研究[J]. 计算机科学, 2013, 40(3): 287-290.
- [15] LONG J. Research on Hybrid Privacy Models and Algorithms for Collaborative Filtering [D]. Guilin: Guangxi Normal University, 2015. (in Chinese)
- 龙军. 面向协同过滤推荐的混合隐私保护技术和算法研究[D]. 桂林: 广西师范大学, 2015.
- [16] MCSHERRY F, MIRONOV I. Differential private recommender system; building privacy into Netflix prize contenders[C]// *Proc. of KDD*. 2009: 627-636
- [17] XIAN Z Z, LI Q L. Research on application of differential privacy in recommender system[J]. *Application Research of Computers*, 2016, 33(5): 1549-1557. (in Chinese)
- 鲜征征, 李启良. 差分隐私保护在推荐系统中的应用研究[J]. 计算机应用研究, 2016, 33(5): 1549-1557.
- (上接第 128 页)
- [3] HUO Z, MENG X F, HUANG Y. PrivateCheckIn: Trajectory Privacy-Preserving for Check-In Services in MSNS[J]. *Chinese Journal of Computers*, 2013, 36(4): 716-726. (in Chinese)
- 霍峥, 孟小峰, 黄毅. PrivateCheckIn: 一种移动社交网络中的轨迹隐私保护方法[J]. 计算机学报, 2013, 36(4): 716-726.
- [4] XU T, CAI Y. Exploring Historical Location Data for Anonymity Preservation in Location-Based Services[C]// *The 27th Conference on Computer Communications*. 2008: 547-555.
- [5] SHI M Y. Research on Trajectory Privacy Protection in Location Based Service[D]. Nanjing: Nanjing University of Posts, 2014. (in Chinese)
- 史敏仪. 面向位置服务的轨迹隐私保护技术研究[D]. 南京: 南京邮电大学, 2014.
- [6] PALANISAMY B, LIU L. Attack-resilient mix-zones over road-networks: architecture and algorithms[J]. *IEEE Transactions on Mobile Computing*, 2015, 14(3): 495-508.
- [7] KIDO H, YANAGISAWA Y, SATOH T. An anonymous communication technique using dummies for location-based services [C] // *International Conference on Pervasive Services*, 2005 (ICPS'05). 2005: 88-97.
- [8] SUZUKI A, IWATA M, ARASE Y, et al. A user location anonymization method for location based services in a real environment [C]// *Sigspatial International Conference on Advances in Geographic Information Systems*. ACM, 2010: 398-401.
- [9] KATO R, IWATA M, HARA T, et al. A dummy-based anonymization method based on user trajectory with pauses[C]// *International Conference on Advances in Geographic Information Systems*. 2012: 249-258.
- [10] KATO R, IWATA M, HARA T, et al. User Location Anonymization Method for Wide Distribution of Dummies[M]// *Database and Expert Systems Applications*. Springer Berlin Heidelberg, 2013: 259-273.
- [11] YOU T H, PENG W C, LEE W C. Protecting Moving Trajectories with Dummies[C]// *International Conference on Mobile Data Management*. IEEE, 2007: 278-282.
- [12] WU X, SUN G. A Novel Dummy-Based Mechanism to Protect Privacy on Trajectories[C]// *IEEE International Conference on Data Mining Workshop*. IEEE, 2014: 1120-1125.
- [13] LEI P R, PENG W C, SU I J, et al. Dummy-Based Schemes for Protecting Movement Trajectories [J]. *Journal of Information Science & Engineering*, 2012, 28(2): 335-350.
- [14] BRINKHOFF T. Generating Network-Based Moving Objects [C]// *International Conference on Scientific and Statistical Database Management*. IEEE Computer Society, 2000: 253-255.
- [15] MONTGOMERY D C, KOWALSKI S M. Design and Analysis of Experiments; Minitab Companion[M]. Wiley, 2010.