隐私保护的共享协同过滤算法研究

目录

[引言 2](#_Toc512267006)

[研究历程与现状 2](#_Toc512267007)

[协同过滤 3](#_Toc512267008)

[向量相似性度量 3](#_Toc512267009)

[基于用户的协同过滤 3](#_Toc512267010)

[基于项目的协同过滤 3](#_Toc512267011)

[局部敏感哈希 3](#_Toc512267012)

[局部敏感哈希的意义 3](#_Toc512267013)

[局部敏感哈希与协同过滤相结合 3](#_Toc512267014)

[同态加密 4](#_Toc512267015)

[同态加密的概念 4](#_Toc512267016)

[Paillier同态加密及其性质 4](#_Toc512267017)

[Paillier同态加密与协同过滤相结合 4](#_Toc512267018)

[参考文献 4](#_Toc512267019)

# 引言

随着互联网和普及和电子商务的发展，推荐系统逐渐成为该领域的一个重要的研究内容，受到越来越多的研究者的关注。如今几乎所有大型电子商务网站都不同程度的提供了各种形式的推荐系统。

协同过滤(Collaborative Filtering, CF)算法是个性化推荐系统中诞生较早，较为成功且应用非常广泛的技术之一。影响协同过滤算法推荐质量的主要因素之一是目标用户的K最近邻用户准确度，算法随后基于这些用户对目标用户尚未评价的项目进行预测。

协同过滤算法依赖用户历史行为数据。随着电子商务系统规模的扩大，用户数目与项目数目急剧上升，并且用户评分数据稀疏性极高，传统协同过滤算法无法保证推荐的实时性。针对这种现象，提出基于局部敏感哈希的协同过滤算法，将评分数据降维并存储在哈希表中，降低近邻用户计算成本。

更大的数据集有助于协同过滤算法得到更准确的推荐结果。但这些数据通常分别属于不同平台，出于数据价值和用户隐私的考虑，这些平台不愿意将数据提供给其他方。针对这种现象，提出基于同态加密的协同过滤算法，保护各平台数据隐私的同时，为用户提供精准的推荐服务。

# 研究历程与现状

数据的隐私保护问题最早由统计学家Dalenius在20世纪70年代末提出，他认为保护数据库中的隐私信息，就是要使任何用户(包括合法用户和潜在的攻击者)在访问数据库的过程中无法获取关于任意个体的确切信息[4]。这具有一定理论上的指导意义，但却是主观和模糊的。以这一定义为参考，学者们在后续的研究中提出了许多明确量化，可操作性强的指标、隐私保护模型和方法。从现有研究来看，k-anonymity[5]及其扩展模型在隐私保护领域具有深远的影响，其基本思想是将数据集里与攻击者背景知识相关的属性定义为准标识符，通过对记录的准标识符值进行泛化和压缩处理，使得所有记录被划分到若干个等价类，每个等价类中的记录具有相同的准标识符值，从而实现将一个记录隐藏在一组记录中，因此这类模型也被称为基于分组的隐私保护模型。然而后续研究表明，这些模型存在两个主要缺陷：首先，这些模型并不能提供足够的安全保障，它们总是因新型攻击的出现而需要不断完善。例如，为了抵制“一致性”攻击，l-diversity[6]、t-closeness[7]、 (a, k)-anonymity[8]、M-invariance[9]等模型相继被提出；文献[10]提出了m-confidentiality模型以抵制“最小性”攻击。“合成式”攻击 [11] ，“前景知识”攻击[12]，“deFinetti”攻击[13]等许多新型的攻击方式都对基于分组的隐私保护模型形成了挑战。出现这一局面的根本原因在于，基于分组的隐私保护模型的安全性与攻击者所掌握的背景知识相关， 而所有可能的背景知识很难被充分定义。因此，与背景知识无关的隐私保护模型逐渐成为研究的热点。其次，这些早期的隐私保护模型无法提供一种有效且严格的方法来证明其隐私保护水平，进而，当模型参数发生改变时，无法对其隐私保护水平进行定量分析，这个缺点削弱了隐私保护处理结果的可靠性。因此，研究人员试图寻求一种新的，鲁棒性足够好的隐私保护模型，能够在攻击者拥有最大背景知识的条件下抵抗各种形式的攻击。差分隐私(Differential Privacy, DP)[14]的提出使得实现这种设想成为可能。差分隐私是Dwork在2006年针对统计数据库的隐私泄露问题提出的一种新的隐私定义，在此定义下，对数据集的计算处理结果对于具体某个记录的变化是不敏感的，单个记录在数据集中或者不在数据集中对计算结果的影响微乎其微。所以，一 个记录因其加入到数据集中所产生的隐私泄露风险被控制在极小的。在可接受的范围内，攻击者无法通过观察计算结果而获取准确的个体信息。 差分隐私能够解决传统隐私保护模型的两个缺陷。首先，差分隐私保护模型假设攻击者能够获得除目标记录外所有其它记录的信息，这些信息的总和可以理解为攻击者所能掌握的最大背景知识，在这一最大背景知识假设下，差分隐私保护无需考虑攻击者所拥有的任何可能的背景知识，因为这些背景知识不可能提供比最大背景知识更丰富的信息；其次，它建立在坚实的数学基础之上，对隐私保护进行了严格的定义并提供了量化评估方法，使得不同参数处理下的数据集所提供的隐私保护水平具有可比较性。因此，差分隐私理论迅速被业界认可，并逐渐成为隐私保护领域的一个研究热点。

# 协同过滤

## 向量相似性度量

协同过滤算法通过相似性来发现用户或项目关系群，相似性主要通过对象的特征向量或属性集合间的相似系数，相关系数以及相异距离等指标度量。常用的度量方式包括欧氏距离，Jaccard相似系数，余弦相似系数，修正的余弦相似系数和皮尔逊相关系数等。这些度量方式分别适用于不同的问题场景。

## 基于用户的协同过滤

基于用户的协同过滤算法基本思想是通过目标用户的K最近邻用户历史数据，加权平均标准化后得到对目标用户未评分的项目的预估，从而给出推荐结果，其中权值为用户间的相似性，相似性根据用户历史数据形式和应用场景选择不同的度量方式。

## 基于项目的协同过滤

通常收集到的用户历史数据所形成的矩阵具有维度高且非常稀疏的特点，在用户量极大的条件下每一位用户仅与众多项目中的极少部分存在关联，因此计算用户特征向量间的相似性的效率较低。并且由于用户的偏好并不稳定，因此推荐结果更新频率需要控制的比较高。出于以上原因，基于项目的协同过滤算法诞生了，其基本思想与基于用户的协同过滤算法类似，只是翻转了用户和项目之间的关系。基于项目的协同过滤算法计算的是项目之间的相似性，由于这种相似性通常更加稳定，因此其计算结果可以被相对较长时间的储存下来，并多次参与推荐过程。

# 局部敏感哈希

## 局部敏感哈希的意义

局部敏感哈希(Locality Sensitive Hashing, LSH)是近似最近邻算法中最流行的之一，具有坚实的理论依据和优异的高维数据空间表现，其主要作用是从海量数据中挖掘近似数据，在文本相似度检测，网页搜索领域有着广泛应用。局部敏感哈希与普通的哈希函数的区别在于，局部敏感哈希能够保留数据的相似性，即如果原数据相似性较高，则他们有很大概率被映射进同一个桶中。局部敏感哈希的本质上是基于概率的算法，因此我们通常引入多张哈希表以获得结果的准确性。

## 局部敏感哈希与协同过滤相结合

对于基本的协同过滤算法，存在数据量较大时的实时性差问题。随着用户和项目数量的上升，用户或项目间两两度量相似度时间成本过高，导致无法实时产生推荐结果。但如果我们运用局部敏感哈希对用户特征向量进行处理，将可能相似的用户较大概率的映射进同一个桶中。在一次针对目标用户的推荐过程中，我们仅需考虑所有哈希表中与其位于同一桶中的元素即可，因此实际参与计算的用户或项目数量大大降低，有效的提升了推荐系统的性能。

# 同态加密

## 同态加密的概念

同态加密(Homomorphic Encryption)是一种加密形式，它允许数据用户对密文进行特定的代数运算，运算得到的结果解密后与对明文进行同样运算的结果一样。换言之，这项技术令人们可以在加密的数据上进行注入检索，比较等操作并得到结果，却无需在处理过程中对数据进行解密，也因而无法得知明文内容。其意义在于，这种技术能够从根本上解决将数据及操作委托给第三方时的安全问题，因此在各种云计算场景中有着广阔的应用场景。

## Paillier同态加密及其性质

Paillier同态加密是基于和数分解的困难性的概率公钥密码系统，是一个满足加法同态的加密体制。若用E表示此加密过程，则Paillier同态加密满足的性质可以具体表述为：

并由此延伸得到另一条常用性质：

通过这两个性质可以帮助我们在仅持有对方密文及公钥情况下得到各自明文的比较结果。

## Paillier同态加密与协同过滤相结合

协同过滤过滤算法推荐结果的准确性依赖于寻找目标用户或项目的最近邻个体的准确性和相似程度，而更大的数据集有助于寻找到更多于目标用户或项目相似度较高的个体。在现实情况中，单一平台的数据量是有限的，一种直接而有效的数据来源是其他平台，但出于信息价值，组织及用户隐私性的考虑，这些数据通常不会在不同平台间共享。因此，需要一种方式在不暴露数据明文的情况下利用其他平台的数据产生综合推荐结果的方式，Paillier同态加密能够在此处发挥作用。

# 参考文献

1. Shen Xin-di, LSH-based Privacy Preserving POI Recommendation Algorithm
2. LIU Shu-shu, Preserving data privacy in social recommendation, 2015, 12
3. Frank McSherry and Ilya Mironov, Differentially Private Recommender Systems, 2009, 7
4. XIONG Ping, ZHU Tian-Qing, WANG Xiao-Feng, A Survey on Differential Privacy and Applications, 2014, 1
5. Sweeney L, k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570
6. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M, l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data(TKDD), 2007, 1(1): 3
7. Li N, Li T, Venkatasubramanian S, t-closeness: Privacy beyond k-anonymity and l-diversity//Proceedings of the IEEE International Conference on Data Engineering(ICDE). Istanbul, Turkey, 2007: 106-115
8. Wong R C-W, Li J, Fu A W-C, Wang K, (a, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA, 2006: 754-759
9. Xiao X, Tao Y, M-invariance: Towards privacy preserving re-publication of dynamic datasets//Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, Beijing, China, 2007: 543-554
10. Wong R C-W, Fu A W-C, Wang K, Pei J, Minimality attack in privacy preserving data publishing//Proceedings of the 33rd International Conference on Very Large Data Bases, Vienna, Austria, 2007: 543-554
11. Ganta S R, Kasiviswanathan S P, Smith A, Composition attacks and auxiliary information in data privacy//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, USA, 2008: 265-273
12. Wong R C-W, Fu A W-C, Wang K, et al, Can the utility of anonymized data be used for privacy breaches? ACM Transactions on Knowledge Discovery from Data(TKDD), 2011, 5(3): 1-24
13. Kifer D, Attacks on privacy and deFinetti’s theorem//Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, Providence, Rhode Island, USA, 2009: 127-138
14. Dwork C, Differential privacy//Proceedings of the 33rd International Colloquium on Automata, Languages and Programing, Venice, Italy, 2006:1-12
15. LI Hong-mei, Collaborative Filtering Recommendation Algorithm Based on Improved Locality-Sensitive Hashing, 2015, 10
16. Gentry, A Fully Homomorphic Encryption Scheme, 2009.9