# 防倒卖设计文档

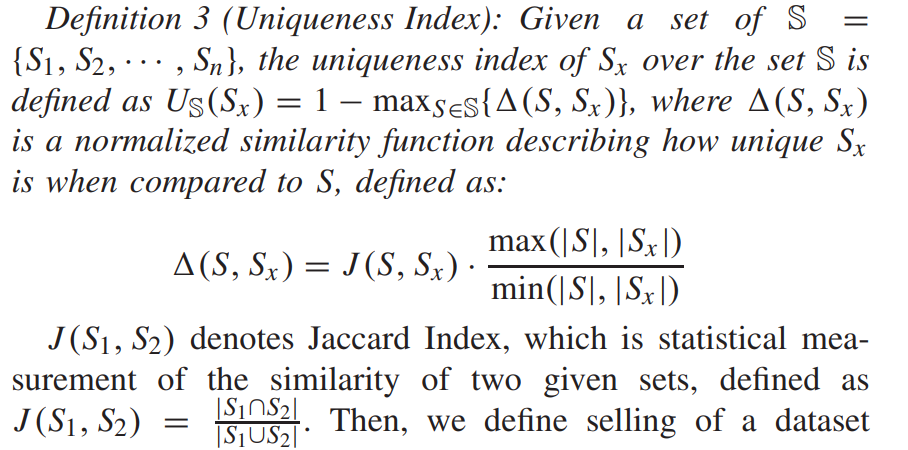
## 业务需求

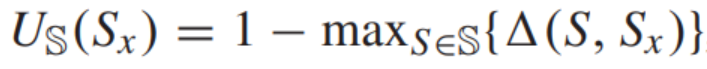
数据倒卖是目前数据交易市场面临的一大痛点。数据由于其特殊性———“所见即所得”，无复制成本，数据转卖成为数据交易中的一大难题。数据存储方、交易平台和数据请求者在获得数据后都可能对数据进行转卖，造成数据所有者利益受损。在集中式的数据交易模式下，一些数据所有者选择对数据加密，这一方法虽然可以避免数据存储方和交易平台对数据进行转卖，但仍然无法规避被数据请求者再次转卖的风险。

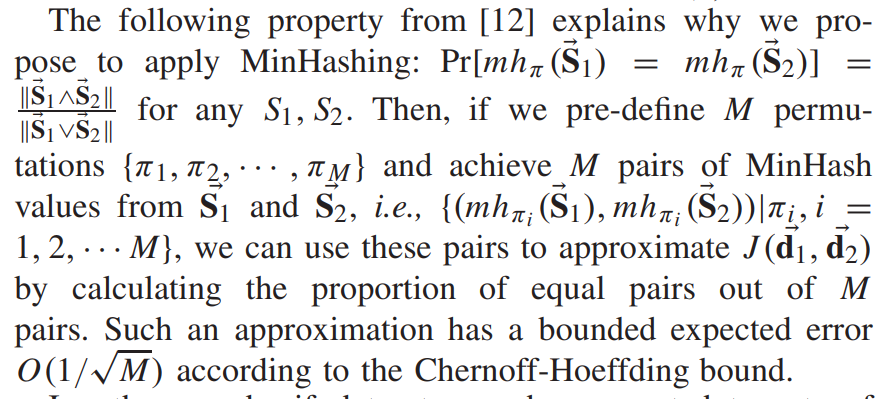
中科大团队在论文*AccountTrade: Accountable Protocols for Big Data Trading Against Dishonest Consumers*中提出了唯一性指数（uniqueness index），该指数可以很好地衡量数据的唯一性，该方法由一系列目前业界最前沿的不同类型数据（视频，图片，表格，文章，JSON）查重算法组成，本文将介绍其原理，并利用唯一指数实现防倒卖。

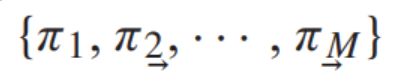
## 基本原理

对于信息提取（IR)技术，文章，图片，视频都被表征为元素的集合。例如，假设原始数据是'bad boy'，利用k-grams算法（k = 4)可以得到集合{“bad ”, “ad b”, “d bo”, “ boy”} 。下图给出了唯一指数的定义：



如上图所示，唯一指数由公式

定义，其中Jaccard index表征两个集合的重合程度（也就是两份数据的重合程度）。计算两个集合的交集时间复杂度为O(n)，论文中给出了一种时间复杂度O(1)的估算方法——借助minHash：

简单来说图中公式的含义为S1的minHash与S2的minHash相等的概率等于jaccard index, 我们估算出前者，就能获得jaccard index的估算值（估算结果可靠并且时间复杂度为O(1))；对于前者，我们需要预先定义M个全排列向量

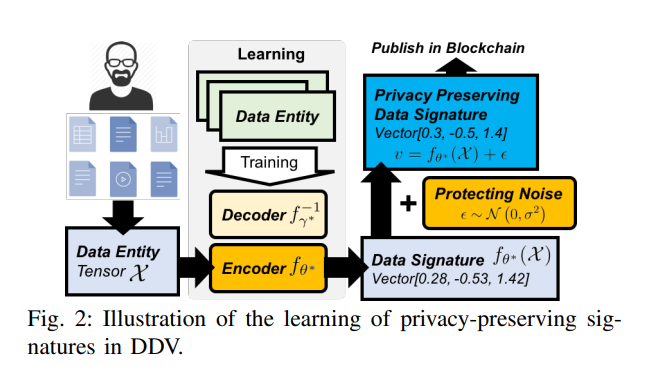
然后计算出M对S1和S2的minHash，最后，用其中相等的除以总量就可以获得jaccard index的估计值，从而获得唯一指数。

对于我们的DDM，我们需要设置一个存储代理。每次卖家上架数据的时候，都需要对代理内的全量数据进行scan，计算数据的唯一指数并检查是否构成转卖，如果合法，则将数据转存进入代理。

## 方案二数据嵌入

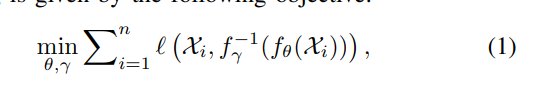
论文*ZHOU J,TANG F,ZHU H,et al．Distributed data vending on blockchain*提出一种结合数据嵌入和距离度量学习的技术，它首先学习一个隐私保护签名来有效地总结原始隐私数据，然后使用通过选择的相似性度量来计算相似性，最后从数据库中实现智能合约的检索，无需损害私有数据的完整性。

## 实现原理

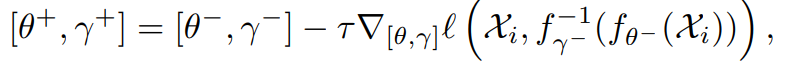


### 数据嵌入

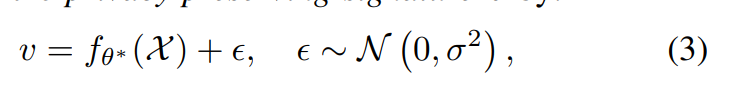
其中θ和γ分别表示加密函数f和其反函数的一系列参数的向量。

论文中给出了计算θ和γ的目标函数

在有一定的数据样本的情况下，可以通过机器学习算出其最优解，即获得了将高维数据映射成向量的函数f，这个向量可以在最大程度上保持原有数据特征，买家将这个一维向量发布到链上。

由于这个方法需要样本量的支撑，在运营初期可能没有那么多的样本供机器学习，所以论文提出了一种随样本量不断优化学习的方法：

另一方面，由于加密函数f是公开的，其反函数f-1也是公开的。在某个买家在购买了一定量的数据之后，可能能够推算出γ从而恢复原始数据，鉴于此，论文做出了一个trade-off，加入了一个高斯噪音来破坏这个可逆的推算过程（然而，这样会影响检索性能）：



### 相似性检索

对于一个查询请求v`, 它和数据嵌入向量v之间的马氏距离决定了它们之间的相似程度：



其中Mt描述了该特定查询的参数矩阵。

用RNN可以轻松实现该算法。

对于我们的DDM，每次卖家发布数据时，都需要将原始数据隐射到低维向量，这个低维向量最大程度地保留了原始数据的特征，并将结果发布到链上，服务器对之前所有的向量进行全量scan，检查它们的距离以判别它们的相似程度。

## 两种方案的优劣

方案一需要设置存储代理，要保留全量数据，这会对数据的隐私性构成挑战，同时对代理的算力要求较高；优点是支持几乎所有数据类型，尤其是表格类型，在这之前还没有对表格类型进行查重的相关研究。方案二不需要设置存储代理，不保留全量数据，隐私性较高，但是不支持表格类型的查重（因为表格的行和列可以任意转换），同时对于其他类型的数据，在样本量不够大的情况下，查重精度远没有方案一高，对算力有一定的要求。优点是简单快捷且隐私性较高。

我们可以在初期样本量不大的时候用方案一，同时对方案二的模型进行训练。等到样本量足够监测精度的时候再使用方案二代替。

## 计划

接下来四周我将分别实现两种算法，计划如下：  
第一周实现方案一，第二周实现方案二，第三周将两种方案结合，第四周测试。