数据资产智能识别方法

王　宁
国家能源信息技术有限公司，中国·北京　100011

【摘　要】本文以元数据权威性及唯一性管理为视角，探讨如何利用一套数据资产智能识别方法，智能识别重复的数据交叉资产，结合动态权重算法，实现对企业数据资产的完整性、权威性、唯一性梳理和管理。帮助推动企业数据资产的整合，持续提升企业数据资产价值。

【关键词】大数据；数据资产；资产目录；智能识别；数据去重

Data Asset Intelligent Identification Method

Wang　Ning
National Energy Information Technology Co., Ltd., China Beijing 100011

[Abstract] From the perspective of metadata authority and uniqueness management, this paper discusses how to use a set of data asset intelligent identification methods to intelligently identify duplicate data cross assets, and combine dynamic weighting algorithms to achieve the integrity and authority of enterprise data assets. , Unique sorting and management. Help promote the integration of enterprise data assets and continuously enhance the value of enterprise data assets.

[Keywords] Big data; Data assets; Asset catalog; Intelligent identification; Data deduplication

随着信息技术不断发展，人类科技迈入了大数据时代，数据资产被认为是这个时代最重要的资产形式之一，而且发挥着越来越重要的作用。

数据资产被认为是组织合法拥有或者控制的、可计量的、能为组织带来价值的数据资源[1]。它的企业资源化是通过建立企业级数据资产目录，将数据像管理物理资产一样进行管理[2]。企业数据资产目录是企业数据资产的“台账信息”、是展示的“窗口”、是企业数据资产价值发挥的“关键所在”[2]。数据使用方通过数据资产目录了解企业数据资产信息，并可根据实际需求申请数据资源信息使用，“所见即所得”的数据资产目录形式，有助于提高数据使用效率、降低数据重复开发成本，是数据发挥资产价值的基础。同时数据资产中包含的数据权威系统、数据管理主责部门等信息，还可以协助识别数据管理责任、解决各方数据问题争议。

在企业数据资源治理的过程当中，笔者研究发现，大多数企业的数据治理还处于开始治理阶段。企业的数据资源缺乏合理规划，加之复杂的数据应用环境、往往数据资产存放混乱，数据质量不佳，不能有效为企业决策提供依据，并且占据了大量的信息基础设施资源。因此，数据资产才更需要被治理，这样才能有效对其管理，最大化发挥企业数据资产价值[3]。本文从元数据权威性和唯一性管理的角度，探讨了如何利用一套智能识别方法，促进企业数据资产的集成，不断提高企业数据资产的价值。

1　问题的提出

1.1 消除冗余，破除孤岛

企业在之前的信息化建设中，建成了大量操作类、管理类、决策类的系统，一方面，在技术和应用初期，由于历史原因，势必会形成不可避免的数据隔离。另一方面，随着时间的推移，企业业务在不断扩大，这些孤岛越来越影响业务系统之间顺畅的数据共享，大量的交叉数据，降低了资源利用率和数据的资产效益，使得破除孤岛极其必要。

1.2 提升质量，重建权威

企业在业务经营过程中，由于各类原因，会导致诸多数据质量问题，如数据不完整、数据标准不一致、编码语义不规范、数据重复等问题，给企业决策带来偏差，造成经济损失。企业数据使用者将会逐渐对数据资产丧失信任，数据资产也必将沦为无价值的“垃圾”。唯有通过对数据资产进行科学管理，以及运用行之有效的智能化识别方法，持续整合资产“版图”，不断提升数据质量，才能使得企业数据资产权威得以重建。

1.3 智能支撑，持续治理

企业在发展的过程中，对数据资产意义认识越来越深，会意识到数据资产需要被治理。但当动手处理时，却发现缺乏有效治理的方法和工具支撑，对于这些数据资产类的持续治理工作，一些企业不知道如何开展。另外一些企业下定决心治理，花费了大量人力去进行治理，费时费力，收效甚微。最后，有些企业即便是阶段性提升了数据资产的质量和权威性，但缺乏持续治理能力，使得一度恢复权威性的数据资产又再度陷入混乱之中。目前迫切需要提供一种智能化的方法来不断帮助企业实现数据资产治理。

2　数据资产识别与分析

2.1 数据资产

数据资产是指由个人或企业拥有或者控制的，能够为企业带来未来经济利益的，以物理或电子的方式记录的数据资源[4]。

1、数据资产的物理属性。数据资产的物理属性是指其存储在于存储介质中，占用物理空间[5]。

2、数据资产的价值属性。一方面，大数据领域所拥有的庞大数据信息，另外一方面，只有在这些信息中提炼出数据元

信息，才能使得海量数据能被萃取，提炼，从而变成数据资产，才有可能变成数据价值。不经处理的海量数据信息，只能白白浪费存储资源，而且企业也无法使用这些信息产生价值 。

2.2数据资产目录

数据资产目录按照相似的图书目录形式对不同的数据资产进行分类，以实现用户对数据资产的快速浏览。数据使用者通过数据资产目录了解企业数据资产信息，并可根据实际需求申请数据资源信息使用，“所见即所得”的数据资产目录形式，有助于提高数据使用效率、降低数据重复开发成本，是数据发挥资产价值的基础。

2.3资产元数据

资产元数据是描述资产数据的数据，它不仅仅表示数据的类型、名称、值等信息，还可进一步提供数据的上下文描述信息，例如数据的所属域、取值范围、数据间的关系等[6]。在企业数据资产管理中，它是一个权威的载体，可以在企业的各个业务部门之间重用。所以，抓住了资产元数据，就能够利用其数据稳定、数量少但影响范围广泛的特点，将各种数据资产进行规范化定义和管理。

2.4资产元数据识别

基于唯一性和权威性校验的数据资产智能“去重”识别方法能够有效支撑企业在海量数据资源的使用过程中很好的去除重复数据。该方法主要采用智能比对、智能分析等算法对企业各业务系统及数据资产平台内的元数据信息进行比对校验，来校验数据资源的唯一性和权威性，将经校验的数据对象进行统一的存储管理，通过对数据资产的盘点和发布，为企业创建和提供更高质量的数据支撑服务，并在整个过程中有效降低数据检测的复杂度，提高数据资产平台的运行效率，建立起可供企业实际使用的高质量数据资产。

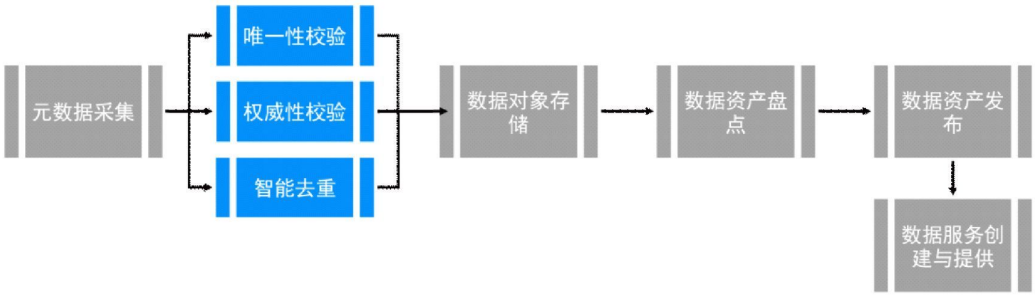

图 1 基于唯一性和权威性校验的数据资产智能去重识别流程

通过对企业数据资产进行唯一性和权威性校验，能够更好的帮助企业在使用数据资源的过程中，对数据资产元数据有效“去重”，大幅提高企业数据资产的质量，为企业发现各项业务潜在的运行特征、规律及趋势，挖掘各业务系统存在的问题与不足，以及预测企业未来业务运营状况和关注重点提供更好的数据支撑。

（1）数据资产的唯一性分析

通过分析传统相似重复记录检测算法的优缺点，总结并建立新型的相似重复检测算法，有效减少数据比较次数，提高算法运行效率和查重准确率，结合关键属性分割和动态权重匹配等算法，将从各业务系统所获取的元数据同数据资产平台既有的元数据进行比对校验，对数据的重复与否打上标签，为后续的删除处理工作提供明确的标识。

（2）数据资产的权威性分析

通过标签识别技术，对数据唯一性分析的成果进行精准定位识别，准确判断出本次采集的数据资产是平台中已有资产的重复数据，并通过人工智能等技术实现对重复数据权威性校验，如定位建立时间较早的数据为权威数据，数据命名规范的数据为权威数据等，确保数据资产来源权威，并对数据再次打上权威标签，对非权威数据做去重处理。

（3）数据资产的智能去重识别

基于数据资产唯一性和权威性的分析结果，采用标签识别技术，精准定位需要去除的重复数据，基于关联属性分割机动态权重匹配等算法，智能的分析和处理重复数据，并执行去重操作，让企业数据资产保持唯一、权威。

3 数据资产智能“去重”方法

数据资产的智能去重识别过程，可归纳为三个步骤，一是元数据预处理，二是唯一性、权威性校验，三是数据资产去重计算及处理，最终实现精准的去重识别和处理。具体如下：

3.1数据唯一性及权威性校验

对数据源采集的元数据信息和数据资产平台已有的元数据信息进行比对，分析数据的重复率，并为数据添加标签，用以标识数据资产的唯一性及权威性。具体如下

3.1.1判断完全重复数据

将获取的元数据与数据资产平台中已有元数据信息进行比对，若数据库表中字段中文名称或字段英文名称完全一样，则判定数据资产完全重复，如下图中的两种情况，其中A表为数据资产平台中的资产信息，B表为从业务系统采集的数据资产信息，若A表中有且仅有的3个字段中文名称或英文名称与B表中有且仅有的3个字段中文名称或英文名称（不区分大小写）完全一样，则视为数据资产完全重复。

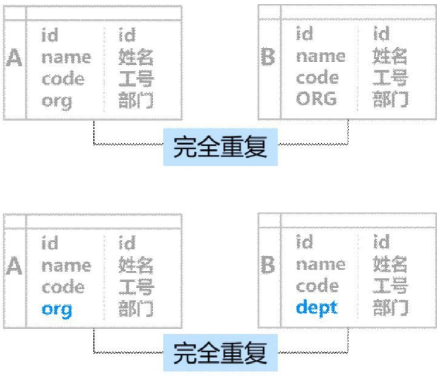

图 2 数据资产完全重复

3.1.2判断部分重复数据

将获取的元数据与数据资产平台中已有元数据信息进行比对，若数据库表中字段中文名称或字段英文名称部分一样，则判定数据资产部分重复。如下图中，其中A表为数据资产平台中的资产信息，B表为从业务系统采集的数据资产信息，A表表名“oep_tsw_zttsw_cqdyqk”，B表表名“tsw_zttsw_cqdyqk”，两表有连续一致的字符“zttsw_cqdyqk”，此时两表可视为数据资产部分重复。

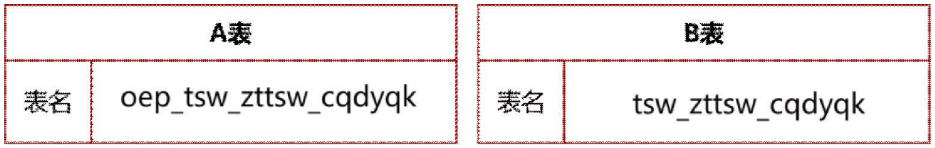

图 3 数据资产部分重复

3.1.3计算数据重复率并添加标签

通过对获取的元数据英文表名或中文表名中出现的连续一致字符的个数与数据资产平台中元数据的英文表名或中文表名个数进行比对计算，从而获得两表之间的表名重复率，用以判定数据资产唯一性。

表名重复率计算公式=采集元数据（英文/中文）表名的连续一致字符个数/数据资产平台元数据（英文/中文）表名总字符个数*100%

数据标签作为数据唯一性及权威性的结果标识，其计算公式采用中文名重复率*权重+英文名重复率*权重计算得出，其中权重的具体占比，可根据数据的实际情况进行适当调整，以提升对数据重复率判断的准确性。

3.2数据资产去重计算

首先需要利用业务领域相关知识，人工选定各数据资产的相关业务属性，依据数据集进行互无交叉的分割，然后借用聚类思想，使用开销较小的算法对数据进行粗聚类，并行在数据集中进行两两比对，精确定位重复数据。其次在数据资产平台上利用基于动态权重的数据重复计算方法，智能识别重复数据，精确检测相似记录，提高数据资产质量，降低数据重复率。

针对数据信息可匹配比较的特点，提出了一种基于动态权重的匹配算法，以提高匹配精度，持续精准调整数据所匹配的权重，确保能够准确识别重复数据资产。

3.2.1关键属性分割

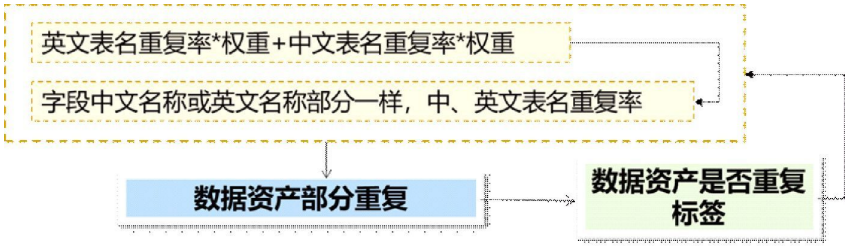

图 4 数据标签建立示意

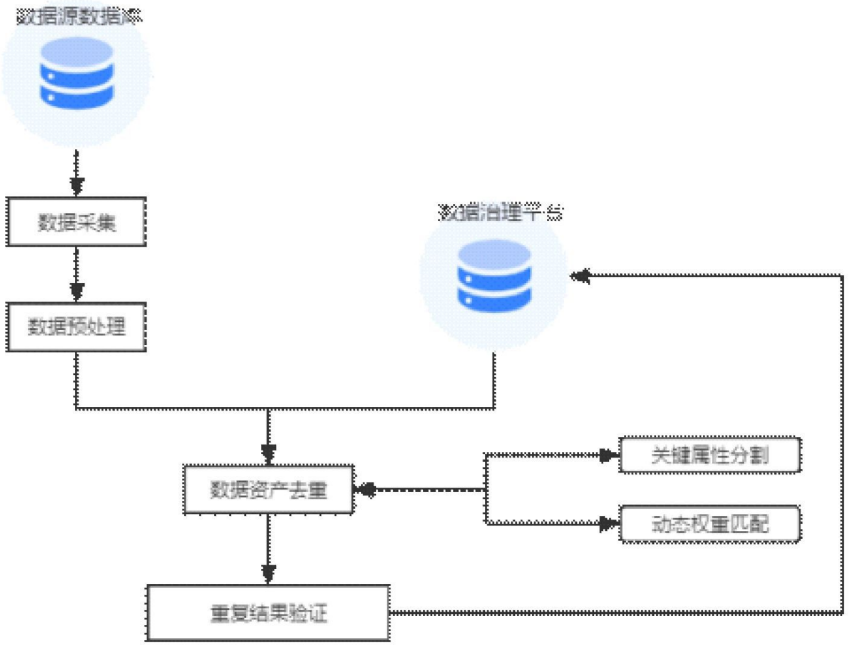

图 5 数据智能识别关系图

一张数据表有若干个属性值组成，属性值描述了数据表具体记录的实体内容，众多属性值中，各项属性的重要性是不同的，至少会有一个关键属性对重复数据匹配起到关键作用。应用该特性帮助用户细分关键业务属性，具体流程可以考虑以下几个方面：

1）第一步，选择关键属性Ka_1，并枚举其所有的M个属性，并将具有相同属性值的元素归约成一组，可将大的数据集划分为M个没有重合的数据小集合。

2）第二步，如果结果不理想，可以选择Ka_2，重新对这些数据集进行划分。

3）第三步，如果结果仍然不理想，可重复第二步，直到划分的数据集满足要求。

2、动态权重匹配算法设计

通过动态权重匹配算法计算数据资产重复率和权重占比，根据计算结果，对现有数据资产打标签，即新采集数据资产与现有资产不重复。

动态权重匹配算法设计上，需要选取计算数据记录之间相似度的方法，虽然这些方法是多种多样，但编辑距离的算法是其中最为简单和使用广泛的，其原理简单易于理解，算法清晰便于操作，本文将采用这种方法来计算动态相似权重。

假设，数据集$D=\{d_1, d_2, d_i\cdots,\}$，$d_i$表示数据集的第i条记录，属性集$C=\{c_1, c_2, \cdots, c_n\}$，$c_k$表示该数据集合n个属性中的第k个属性，$d_{ik}$表示数据$d_i$条记录的第k个属性值，那么对于任意两条数据$d_i$和$d_j$在第k个属性上的值分别为$d_{ik}$和$d_{jk}$，令$L(d_{ik}, d_{jk})$表示其编辑距离，相似度为$Sim(d_{ik}, d_{jk})$可以用以下公式来表示：

$$Sim(d_{ik}, d_{jk})=1-\frac{L(d_{ik},d_{jk})}{\max(|d_{ik}|,|d_{jk}|)}$$

$L(d_{ik}, d_{jk})\neq\infty$

$Sim(d_{ik}, d_{jk})=0$　　$L(d_{ik}, d_{jk})=\infty$

为了方便计算和操作，需要计算的数据间相似度，在此可以视为所有与之对应属性相似度的加权平均值，故假设各属性的贡献比例为：v_1：$v_2\cdots:v_n$，那么这些属性的总权值为1，用W_k计算属性C_k权重的计算公式为：

$$W_k=v_k/\sum v_i$$

任意两条数据的重复度可以用以下公式来表示;

$$Sim(d_i,d_j)=\sum_{k=1}^{n}Sim(d_{ik},d_{jk})*W_k$$

由于此算法在计算过程中，若数据的某条属性信息缺失，则该属性相似度为0，这样数据重复度的总值会下降，两条数据将被视为不同数据。然而d_i, d_j有可能是重复的数据，在这种情况下，会给重复数据识别带来了误差，所以本文提出了一种改进方法，应对属性失效的情况。

假设 Val_k 表示属性k的有效性，则

Val_k =1，属性k有效；

Val_k =0，属性k无效。

倘若某个属性不符合要求，则

W_k可以表示为 $W_k-(Val_k*v_k)/\Sigma(Val_i*v_i)$ ，这样一来，整个有效数据的属性总的权重1，避免了属性不完整情况下计算的影响。所以，基于动态权重数据间的重复度计算公式为：

$$Sim(d_i,d_j)=\sum_{k=1}^{n}Sim(d_{ik},d_{jk})*\frac{Val_k*W_k}{\Sigma_{i=1}^{n}Val_i*W_i}$$

4　结论

本文设计了在数据资产使用过程，如何对不具备唯一性和权威性的数据进行智能识别和去重操作，通过本方法的数据预处理、数据唯一性及权威性判断比对、以及标签标识和识别等技术，实现数据的智能去重，在数据去重计算过程中，还提出了关键的动态权重匹配等算法，该项技术通过对数据重复权重的持续智能调整，能够更精准的分析和判断企业数据资产的重复率，提高集团数据资产平台的数据质量，为企业数据战略的发展提供可靠的支撑和保障。

参考文献：

[1]全国信息技术标准化技术委员会.信息技术服务治理 第5部分: 数据治理规范:GB/T34960.5-2018[S].2018.

[2]郑高峰,秦丹丹,刘丽,等.基于知识图谱技术的数据资产管理设计与应用验证研究[J].中国科技投资,2020,(7):61-63.

[3]张宁.主数据驱动视角下的企业档案数据资产管理[J].档案学研究,2019(6):47-52. DOI:10.16065/j.cnki.issn1002-1620.2019.06.007.

[4]赵璐.数据资产评估过程难点分析及建议[J]. 全国流通经济,2021(21):131-134. DOI:10.3969/j.issn.1009-5292.2021.21.041.

[5]朱扬勇,叶雅珍. 从数据的属性看数据资产[J]. 大数据,2018,4(6):65-76. DOI:10.11959/j.issn.2096-0271.2018062.

[6]单德祥.基于XMI的元数据交换技术研究及其应用[D].北京:北京邮电大学,2006.

作者简介：

王宁（1985.12-），女，汉族，河南商丘人，学士，高级咨询顾问，从事数据分析、数据挖掘。