Hans Journal of Data Mining 数据挖掘, 2022, 12(1), 80-89
Published Online January 2022 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2022.121009

Hans汉斯

建立多源异构编码映射关系的实践

谢　芬[1]，王　敏[2]，陈方圆[3]

[1]云赛智联股份有限公司，上海
[2]星环信息科技(上海)股份有限公司，上海
[3]上海海事大学，上海

收稿日期：2021年12月19日；录用日期：2022年1月19日；发布日期：2022年1月27日

摘　要

在政务大数据中心的数据治理过程中，不同政务场景下，由于管理策略不同，业务过程对实体对象的相同属性信息的记录会有不同的数据编码结构，这就是政务数据多源融合过程中常见的一种难题。与常规方法不同，本文通过引入统计学的列联相关分析法，解决了不同业务场景的异构法人登记属性融合问题，建立了准确的映射关系。此实践将统计学方法应用到多源异构政务数据融合过程中，不仅快速、低成本的解决了实际问题，并且对于解决其他数据融合问题具有较高的参考价值。

关键词

政务大数据，数据治理，数据融合，异构编码，编码映射，列联分析

Practice of Establishing Multi-Source Heterogeneous Coding Mapping Relationship

Fen Xie[1], Min Wang[2], Fangyuan Chen[3]

[1]Inesa Intelligent Tech Inc., Shanghai
[2]Transwarp, Shanghai
[3]Shanghai Maritime University, Shanghai

Received: Dec. 19th, 2021; accepted: Jan. 19th, 2022; published: Jan. 27th, 2022

Abstract

In the data governance process of government big data center, under different government scenarios, due to different management strategies, the records of the same attribute information of entity

文章引用: 谢芬, 王敏, 陈方圆. 建立多源异构编码映射关系的实践[J]. 数据挖掘, 2022, 12(1): 80-89.
DOI: 10.12677/hjdm.2022.121009

objects in business processes will have different data coding structures, which is a common problem in the process of multi-source fusion of government data. Different from the conventional methods, this paper solves the problem of heterogeneous legal person registration attribute fusion in different business scenarios by introducing the statistical column correlation analysis method, and establishing an accurate mapping relationship. This practice applies statistical methods to the process of multi-source heterogeneous government data fusion, which not only solves practical problems quickly and at low cost, but also has the high reference value for solving other data fusion problems.

Keywords

Government Big Data, Data Governance, Data Fusion, Heterogeneous Coding, Coding Mapping, Contingency Analysis

Copyright © 2022 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

根据《政务信息资源共享管理暂行办法》(国办发〔2017〕39 号)要求，要加快推动政务信息系统互联和公共数据共享，充分发挥政务信息资源共享在深化改革、转变职能、创新管理中的重要作用，增强政府公信力，提高行政效率，提升服务水平[1]。目前政务大数据中心的建设日趋完善，政府部门数据共享也在积极推动，打破信息壁垒，优化政府管理流程和提升协同治理能力成为当务之急[2]。随着数据共享交换逐步深入，对数据质量的要求也在逐步提升，出现了大量的多源异构政务数据融合的数据治理需求。本文就政务数据多源融合过程中常遇到的编码映射问题，通过尝试统计学的方法，为不同业务场景下产生的异构法人登记状态建立起准确的映射关系。

2. 现状与问题

政务管理工作体系庞大而复杂，相同的实体对象，例如自然人、法人等，在不同的政务场景下都有数据产生。不同的管理层次、管理策略，加上辅助政务的系统由不同的开发商在不同的年代开发，因此对实体数据的记录方式也有着较大的差异，进而产生了大量的政务异构数据。

与此同时，不同政务场景下，不同的政务系统对实体的相同属性建立了不同的策略的数据编码规则，这些编码都是根据场景的需求及当时的管理策略建立的，日积月累而成，因此不同的场景下的相同实体属性编码存在较大的差异[3]。

政务大数据中心为了满足不同政务部门的数据应用需求，提供质量更高的数据服务，就需要想方设法融合这些多源异构编码，即建立编码映射关系。

常规的方法基本上都是依靠数据来源政务部门的协调来寻求编码的一致性或建立映射关系，但这种方法周期长，难度大，并且难以持续应对实际政务过程中的不断变化[4]。

因此，在不断追求数据融合质量、效率的目标引导下，政务大数据中心不断探索、寻求更优化的方法来解决多源异构编码融合问题。本文通过一个实际问题的实践，探索运用统计学算法来解决此类问题。

3. 异构编码融合问题实例

在政务大数据中心法人实体数据治理过程中，出现了如下的实际问题。

法人实体(包含法人和其他组织，以统一社会信用编码的管理范畴为准)的数据有两个来源 A、B，所

包含的法人数据范围有部分重叠，例如都有企业、个体工商户等类别法人实体，同时也都有另一来源没有的类别法人实体。

根据数据融合需求，融合后法人实体数据使用 A 来源法人实体登记状态编码，因此需要建立 B 来源编码和 A 来源编码的映射关系。

通过数据探查发现两个来源的法人实体登记状态有着不同的定义和编码结构，具体的情况如下：

Table 1. A source legal entity registration status code
表 1. A 来源法人实体登记状态编码表

A 来源法人实体登记状态	
编码	**编码描述**
0001	确立
0002	吊销
0003	注销
0004	撤销
0005	迁往外省市

Table 2. B source legal entity registration status code
表 2. B 来源法人实体登记状态编码表

B 来源法人实体登记状态		
编码	**编码描述**	**说明**
0001	确立	
0002	迁移	
0003	停业	
0004	注销	
0006	吊销未注销	
0008	吊销已注销	针对企业类实体使用
0009	条线变更	
0010	吊销已注销	针对个体工商户类实体使用
0011	迁往外省市	
0012	撤销	

从表 1、表 2 中可以发现两种编码的逻辑规则有较大差异，并且发现如下的问题：

1) 一码多义，在 A 来源的编码中，“0003”的含义是“注销”，而在 B 来源的编码中，“0003”的含义是“停业”；

2) 一义多码，在 A 来源的编码中，“注销”的编码是“0003”，而在 B 来源中“注销”的编码是“0004”；

3) 含义模糊，B 来源中，“0002，迁移”、“0007、迁出”，其含义无法确定是指迁出区县、还是迁出城市。

通过对比分析，发现两个来源的相同法人实体登记状态不同，如表 3 所示，A 来源中状态为“确立”的法人实体，在 B 来源中对应了多种编码，同样，B 来源中状态为“确立”的法人实体，在 A 来源中也对应了多种编码。

Table 3. A and B source legal entity registration status difference statistical sample data
表 3. A、B 来源法人实体登记状态差异统计样例数据表

A 来源		B 来源		数据量(条)
编码	编码描述	编码	编码描述	
0001	确立	0002	迁移	2569
0001	确立	0004	注销	5385
0001	确立	0006	吊销未注销	12,396
0001	确立	0008	吊销已注销	605
0001	确立	0009	条线变更	54
0001	确立	0010	吊销已注销	63
0001	确立	0011	迁往外省市	21
0003	注销	0001	确立	1
0002	吊销	0001	确立	6463
0004	撤销	0001	确立	1
0005	迁往外省市	0001	确立	33
0003	注销	0001	确立	15,341

4. 基于列联分析构建映射关系

目前关于政务数据方面的异构编码的映射问题，相关的文献介绍较少，而且运用统计学的方法来解决政务数据的实际问题更是少之又少。本文通过多方面尝试，经过分析发现列联相关分析法的适用场景和本文所面对的问题非常接近。

当研究两个属性变量之间是否有联系时，需要用到列联表[5]。列联表又称交互分类表，所谓交互分类，是指同时依据两个变量的值，将所研究的个案分类。交互分类的目的是将两变量分组，然后比较各组的分布状况，以寻找变量间的关系。

列联表分析主要包括两个基本任务：一是根据收集的样本数据，产生二维或多维交叉列联表；二是在交叉列联表的基础上，对两个变量间是否存在相关性进行检验。通常情况下，在获得列联表数据之后，我们将会通过统计假设检验两个属性变量是否具有独立性，进而进行列联表分析。

根据算法要求，首先构建列联表，在对列联表行列属性各类别进行命名时，如果法人状态为确立，对应的编码为 0001，则命名为：1-确立，以此类推。最终我们根据两个来源的编码及其对应的编码数量构建的列联表如表 4。

4.1. 列联分析：独立性检验

对于两个分类变量的分析，称为独立性检验，分析过程可以通过列联表的方式呈现，也把这种分析称为列联分析。独立性检验就是分析列联表中行变量和列变量是否相互独立，也就是 A 来源的法人实体登记状态和 B 来源之间是否存在依赖关系。

H_0：两者之间是独立的(不存在依赖关系)

H_1：两者之间不独立(存在依赖关系)

用 SPSS 进行列联分析的独立性检验，输出结果如表 5：

Table 4. A and B source legal entity registration status contingency
表 4. A、B 来源法人实体登记状态列联表

		A 来源					
		1-确立	2-吊销	3-注销	4-撤销	5-迁往外省市	总计
B 来源	1-确立	387,211	6463	15,342	1	33	409,050
	2-迁移	2569	7	24	0	1	2601
	3-停业	0	0	0	134	0	134
	4-注销	5385	0	153,630	0	0	159,015
	6-吊销未注销	12,396	91,258	563	22	0	104,239
	8-吊销已注销	605	62	0	0	0	667
	9-条线变更	54	1	0	0	0	55
	10-吊销已注销	63	644	8972	0	0	9679
	11-迁往外省市	21	1	4	0	740	766
	12-撤销	1	5	0	281	0	287
	总计	408,305	98,441	178,535	438	774	686,493

Table 5. Chi-square test results
表 5. 卡方检验结果

卡方检验			
	值	自由度	渐进显著性(双侧)
皮尔逊卡方	2,376,277.059[a]	36	0.000
似然比	970,052.472	36	0.000
线性关联	299,352.278	1	0.000
有效个案数	686,493		

a. 12 个单元格(24.0%)的期望计数小于 5。最小期望计数为 0.04。

从表 6 可得，卡方值为 2,376,277.06，相伴概率小于 0.001，故应拒绝原假设，认为两者之间不是独立的，两个来源的数据之间存在依赖关系。

4.2. 列联表中的相关测量

接下来需要探讨的问题是，如果变量之间存在联系，它们之间的相关程度有多大？

对两个变量之间相关程度的测定，主要用相关系数来表示。经常用到的品质相关系数有以下几种。

1) φ 相关系数，是描述 2 × 2 列联表相关程度最常用的一种相关系数。

$$\varphi = \sqrt{\frac{\chi^2}{n}}$$

χ^2 是之前计算出来的，n 为列联表中的总频数，也即样本量。

2) 列联相关系数，又称列联系数，简称 c 系数，主要用于列联表大于 2 × 2 的情况。c 系数的计算公式为：

$$c = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

当列联表中的两个变量相互独立时，系数 $c = 0$，并且它不可能大于 1，这一点从式中也可以看出来。由于其计算简便，且对总体的分布没有任何要求，所以列联表系数不失为一种适应性较广的测度值。

3) V 相关系数，鉴于 φ 系数无上限，c 系数小于 1 的情况，格莱姆提出了 V 相关系数。V 相关系数的计算公式为：

$$V = \sqrt{\frac{\chi^2}{n \times \min\left[(R-1),(C-1)\right]}}$$

当两个变量相互独立时，$V = 0$；当两个变量完全相关时，$V = 1$。所以 V 的取值在 0~1 之间。

根据这些计算公式，用 SPSS 来计算列联相关系数，结果如下：

Table 6. Contingency correlation coefficient
表 6. 列联相关系数

对称测量			
		值	渐进显著性
名义到名义	Phi	1.861	0.000
	克莱姆 V	0.930	0.000
	列联系数	0.881	0.000
有效个案数		686493	

对于同一数据，系数 φ、c，V 的结果不同，此处只看 c，V 的结果，从表 7 可以看出，$c = 0.881$，$V = 0.930$，这两个数值相对来说都是很大的，说明 A 来源的法人状态和 B 来源的法人状态的数据具有很强的相关度。

4.3. 列联表的对应分析

在得到 A、B 两来源之间有很强的相关性，继而研究两来源登记状态编码之间的映射关系。运用对应分析方法就能解决这个问题。它们之间的距离越近，表示它们有差不多一样大的“得分”，从而认为它们相互对应。

SPSS 软件的 Correspondence Analysis 模块是专门进行对应分析的模块，下面是用此模块对本文要解决的问题进行分析，可以得到输出结果如下：

Table 7. Correspondence table
表 7. 对应表

对应表						
B 来源	A 来源					
	1	2	3	4	5	活动边际
1	387,211	6463	15,342	1	33	409,050
2	2569	7	24	0	1	2601
3	0	0	0	134	0	134
4	5385	0	153,630	0	0	159,015
5	0	0	0	0	0	0
6	12,396	91,258	563	22	0	104,239

Continued

7	0	0	0	0	0	0
8	605	62	0	0	0	667
9	54	1	0	0	0	55
10	63	644	8972	0	0	9679
11	21	1	4	0	740	766
12	1	5	0	281	0	287
活动边际	408,305	98,441	178,535	438	774	686,493

Table 8. Summary table
表 8. 摘要表

摘要						
维	奇异值	惯量	惯量比例		置信度奇异值	
			占	累积	标准差	相关性
						2
1	1.145	1.312	0.441	0.441	0.000	−0.131
2	0.885	0.783	0.263	0.704	0.000	
3	0.733	0.538	0.181	0.885		
4	0.586	0.343	0.115	1.000		
总计		2.975	1.000	1.000		

Table 9. Row point overview table
表 9. 行点总览表

行点总览[a]									
B来源	数量	维得分		惯量	贡献				
		1	2		点对维的惯量		维对点的惯量		
					1	2	1	2	总计
1	0.083	−1.716	0.099	0.291	0.214	0.001	0.967	0.002	0.969
2	0.083	−1.809	0.153	0.323	0.238	0.002	0.967	0.005	0.972
3	0.083	0.723	1.738	0.333	0.038	0.284	0.150	0.668	0.818
4	0.083	0.455	−1.538	0.306	0.015	0.223	0.064	0.570	0.635
5	0.083	.	.	.	.	.	.	.	.
6	0.083	−0.031	−0.129	0.242	0.000	0.002	0.000	0.005	0.005
7	0.083	.	.	.	.	.	.	.	.
8	0.083	−1.647	0.140	0.263	0.197	0.002	0.984	0.005	0.990
9	0.083	−1.800	0.164	0.318	0.236	0.003	0.971	0.006	0.978
10	0.083	0.498	−1.491	0.277	0.018	0.209	0.086	0.593	0.678
11	0.083	0.309	−0.148	0.306	0.007	0.002	0.030	0.005	0.035
12	0.083	0.705	1.699	0.316	0.036	0.272	0.150	0.673	0.823
活动总计	1.000			2.975	1.000	1.000			

a. 对称正态化。

Table 10. Column point overview table
表 10. 列点总览表

列点总览[a]									
A来源	数量	维得分		惯量	贡献				
		1	2		点对维的惯量		维对点的惯量		
					1	2	1	2	总计
1	0.200	−2.104	0.151	1.062	0.773	0.005	0.955	0.004	0.959
2	0.200	0.241	−0.142	0.343	0.010	0.005	0.039	0.010	0.049
3	0.200	0.613	−1.414	0.623	0.066	0.452	0.138	0.568	0.706
4	0.200	0.828	1.537	0.686	0.120	0.534	0.229	0.609	0.838
5	0.200	0.422	−0.132	0.428	0.031	0.004	0.095	0.007	0.103
活动总计	1.000			2.975	1.000	1.000			

a. 对称正态化

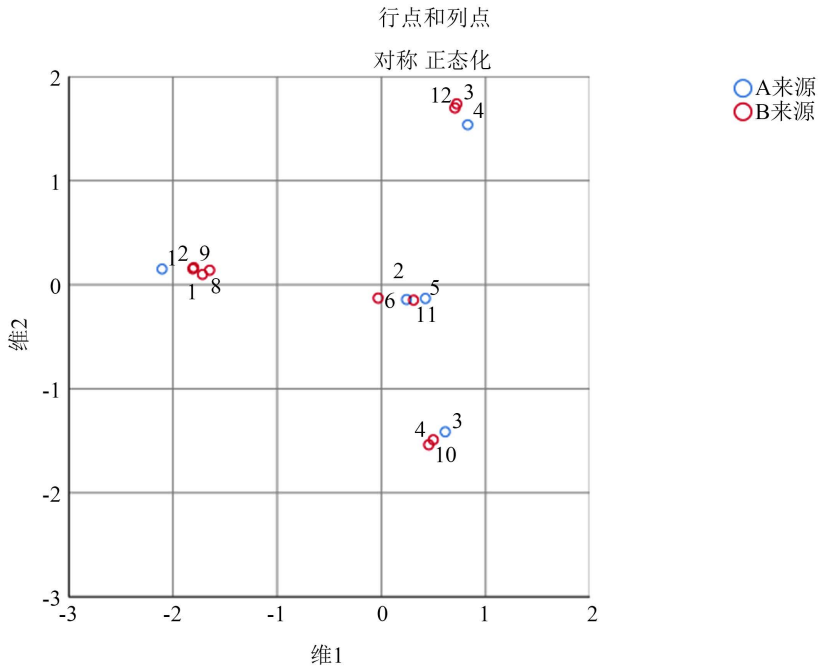

Figure 1. Correspondence analysis diagram
图 1. 对应分析图

其中，输出的第一部分如表 7 对应表是由原始数据按 A 来源与 B 来源的列联表，可以看到观测总数 n = 686493。

第二部分表 8 摘要表给出了总惯量、奇异值及每一维度(公共因子)所解释的总惯量的百分比的信息。奇异值反映的是行与列各状态在二维图中分值的相关程度，惯量比例部分是各维度分别解释总惯量的比例及累计百分比。

第三部分(表 9)和第四部分(表 10)是对列联表行与列各状态有关信息的概括。其中，数量部分分别指列联表中行与列的边缘概率，维得分是各维度的分值，也就是行与列各状态在二维图中的坐标值，惯量

是每一行(列)与其重心的加权距离的平方，贡献部分是指行(列)的每一状态对每一维度(公共因子)特征根的贡献及每一维度对行(列)各个状态的特征根的贡献。由此可以更好地理解维度的来源及意义，如第一维度中，A 来源的 1 类型对应的数值最大，为 0.773，说明 A 来源的 1 状态对第一维度的贡献最大。而对于 A 来源的 2 这一状态，两个维度对其贡献度都不是特别大。

输出的最后一部分图 1 对应分析图是 A、B 来源的各状态同时在一张二维图上的投影。在图上既可以看到每一变量内部各状态之间的相互关系，又可以同时考察两变量之间的相关关系。可以看出，A 来源的五个状态分布较为分散，很明显形成五大类；对于 B 来源，1、2、8、9 被分为一类，3、12 被分为一类，4、10 被分为一类，6 和 11 各自被单独分为一类。同时考察两变量各状态，可以看到 A 来源的 1 状态和 B 来源的 1、2、8、9 状态距离较近，A 来源的 4 状态与 B 来源的 3、12 状态距离较近，A 来源的 5 状态和 B 来源的 11 状态距离最近，B 来源的 6 状态距离 A 来源的 2 状态距离最近，A 来源的 3 状态和 B 来源的 4、10 状态距离较近。

对此可以绘制对应表格如下表 11 所示：

Table 11. Legal entity registration status code mapping table
表 11. 法人实体登记状态编码映射表

A 来源		B 来源	
编码	编码描述	编码	编码描述
0001	确立	0001	确立
		0002	迁移
		0009	条线变更
0002	吊销	0008	吊销已注销
		0006	吊销未注销
0003	注销	0004	注销
		0010	吊销已注销
0004	撤销	0012	撤销
		0003	停业
0005	迁往外省市	0011	迁往外省市

4.4. 映射结果校验

为了验证由此种方法得到的关于异构编码映射关系的准确性，我们利用互联网渠道和信息服务平台，如企查查、天眼查等，采用随机抽样和分层抽样的方法来调查和验证法人实体登记状态数据的拟合度。选取了约 400 个法人实体，通过互联网渠道和信息服务平台查询核对法人实体登记状态，与通过列联分析法融合后的法人实体登记状态相比，最终的重合度达到 96.67%。说明我们构建的映射关系非常准确。

5. 实践小结

本文从政务大数据融合过程中遇到的异构编码问题入手，致力于解决当前典型的异构法人实体登记状态编码映射的难题，通过运用列联相关分析、对应分析等统计学方法，科学、快速、低成本地解决了异构法人实体登记状态编码映射问题，为各部门之间信息共享和业务协同的基础之一，肩负着减少社会负担，提升行政效率的使命，也是建立社会信用体系、企业公信力的重要数据依靠；而且通过第三方公共来源数据验证了该结果的准确性，总体取得了较好的融合效果。

在政务多源异构数据治理融合工作中，类似的情况还有很多，本次实践运用了统计学的部分算法解决了实际问题，为常规的政务多源异构数据治理提供了一种新的思路、方法和参考[6]。以此类推，人工智能、深度学习等新兴技术，也可以运用到数据治理过程中的某些场景，来帮助数据治理人员解决实际问题[7]。

在当前大数据技术高速发展时代，政务数据治理工作有必要打开眼界，运用多学科技术不断降低数据治理成本、提升数据治理效率与质量。

参考文献

[1] 叶战备. 政务数据治理的现实推进及其协同逻辑——以 N 市为例[J]. 中国行政管理, 2021(6): 44-49.

[2] 夏义堃. 政府数据治理的国际经验与启示[J]. 信息资源管理学报, 2018(3): 64-72+101.

[3] 金振坤. 网络编码中优化问题研究[D]: [博士学位论文]. 武汉: 华中科技大学, 2018.

[4] 衡容, 贾开. 数字经济推动政府治理变革: 外在挑战、内在原因与制度创新[J]. 电子政务, 2020(6): 55-62.

[5] 王静龙, 梁小筠. 定性数据分析[M]. 上海: 华东师范大学出版社, 2005.

[6] 曾俊. 大数据驱动“互联网+政务服务”模式创新研究[J]. 中国管理信息化, 2019(8): 161-162.

[7] 沈王恒. 浦东新区政务数据融合服务平台的探索[J]. 信息技术与标准化, 2021(6): 13-18.