

Research and Exploration on the Reliability of Four Character Detection of Electronic Archives Using AI—Taking Document Archives as an Example

Xiangzhen Lin Qingfei Zhou* Panju Hou Yuehao Chen

China Great Wall Technology Group Co., Ltd., Beijing, 100080, China

Abstract

The four character detection is one of the core aspects of the full lifecycle management of electronic archives. The research on the credibility of the four character detection of electronic archives requires comprehensive evaluation from multiple dimensions such as authenticity, integrity, security, availability, predictability, and controllability. Combined with the rapidly developing artificial intelligence model algorithm of archives, the author believes that through the four character detection indicators and AI algorithms, the credibility and credibility of electronic archive management can be given the closest measurement value to reality. The paper explores the use of AI technology to propose a trustworthy evaluation model for measuring the trustworthiness of electronic archives based on four character detection.

Keywords

four character detection; trusted management; full lifecycle; security guarantee

利用 AI 对电子档案四性检测可信度研究探讨——以文书档案为例

林祥振 周庆飞* 侯攀举 陈阅浩

中电长城科技有限公司, 中国·北京 100080

摘要

四性检测是电子档案全生命周期管理工作中核心之一, 电子档案四性检测可信度研究需要从电子档案的真实性、完整性、安全性、可用性、可预测性、可控性等多维度综合评估, 结合档案快速发展的人工智能大模型算法, 笔者认为通过四性检测指标、AI算法, 能够给电子档案管理的可信性和可信度一个最贴近真实的度量值。论文就利用AI技术基于四性检测对电子档案的可信度量方法提出可信评价模型进行研究探讨。

关键词

四性检测; 可信管理; 全生命周期; 安全保障

1 引言

近年来业界对电子档案四性检测已经开展了诸多研究, 2018 年国家档案局发布 DA/T 70—2018《文书类电子档案检测一般要求》^[1], 目前文书档案的四性检测工具软件的研发主要以此为依据, 细化每个阶段的检测项目、检测对象、检测依据和方法, 对电子档案四性检测, 以确保电子档案在

流转过程中是可信的。目前业内的普遍认知是电子档案通过了四性检测中设定的检验指标即为可信任, 但是对于“可信电子档案”研究较少, 即没有一个明确的度量方法或数值来认定一份电子档案在通过网络传输、数据流转的过程中可信度是多少。

2 可信电子档案定义

结合可信计算概念、可信电子文件以及可信电子档案的相关研究, 论文对可信电子档案定义如下: “在电子档案生命周期内, 完成规定的管理要求和业务处理的能力, 这种行为与结果是可以被预期的。”

电子档案可信管理具有以下特征:

①可信不等于安全: 可信强调的是行为结果可预期, 但并不等于行为是安全的, 这是两个不同概念。

【作者简介】林祥振(1974-), 男, 中国河南杞县人, 本科, 从事电子档案管理及理论研究、档案信息化研究。

【通讯作者】周庆飞(1986-), 男, 中国北京人, 高级工程师, 从事国产计算、信息安全和行业数字化转型、电子档案可信管理研究。

②可信具有容错性：可信允许适度的错误发生，这种错误发生是在人们可以接受的综合错误容忍范围内，不影响整体结果并可预期。

③可信具有传递性：可信管理是一个过程，通常一个实体在向既定目标发展的过程中传递预期的结果。

④信任是可度量的：度量必须有一个起点，这个起点必须是绝对可信的，它被叫作度量可信根。

3 评价模型创建思想

可信电子档案评价模型以电子档案数据文件为核心，以档案管理全生命周期各业务环节为主线^[2]，以工作流程为

辅助，围绕电子档案全生命周期内管理过程中所涉及的基础数据、规章制度、业务流程、机构人员和基础技术措施等相关的决策、计划、组织和控制活动开展可信评价研究；主要运用基于四性检测所延伸的可信管理原理、方法和手段，分析和研究各种对电子档案可信造成影响的行为因素，从技术上、组织上和管理上采取有力的措施，解决和消除各种非可信因素，防止各种风险的发生。可信评价模型创建思想如图1所示。

图1体现了电子档案流转的业务主线和日常管理审批工作流程两个方面，这也是电子档案管理的基础。

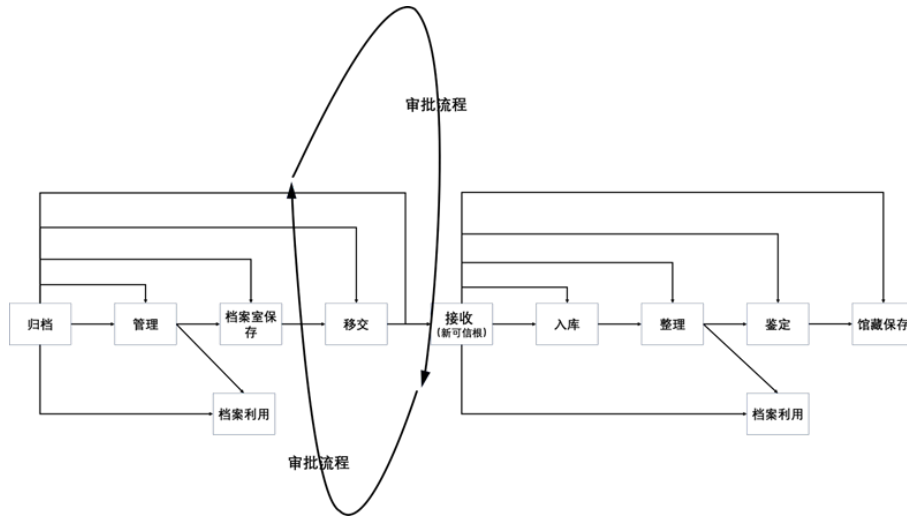


图1 可信电子档案管理模型

4 可信评价模型

4.1 深度学习法——事前分析

众所周知，电子档案管理的业务链从大的环节上讲分为收、管、存、用^[3]，在此基础上又可划分N个更细的中间环节（如移交、接收、归档等），这些环节都是在业务链上所需经过的路径节点，每经过一个节点，可将此作为电子档案本身在此处的一个驻点，我们采用深度学习法对此驻点上所能够涉及的各类行为活动为分析对象，对各类行为活动进行定性和定量化的评估。

深度学习法需要对某一驻点的行为多次分析或对多驻点的同一行为分析才能得出正确的结论，像人工智能一样，需要不断地积累学习，形成学习库，样本越多，时间越长，得到的结论越接近实际。

参考《电子文件归档与管理规范》《电子档案移交与接收办法》《版式电子文件长期保存格式需求》《文书类电子档案检测一般要求》等标准规范对业务、技术及管理行为的要求，在此基础上分析其管理行为并设定指标值。在本方法中，每一个管理行为项值都有其最大值和评估值，其值的来源是一个假定值，也可看作是经验值。最大值的设定可参考等级保护中的安全等级设置来确定，分为1~5级，评估值可以根据用户经验或在某个节点多次实验的结果取值。指

标值越大表示风险影响小，指标值越小表示风险影响越大，用户可根据单个测评指标值进行单项改进，也可根据整体测评评价整个系统的可信性，进一步提出改进策略^[1-4]。

4.2 风险矩阵法——事后评估

风险矩阵法是一种能够把危险发生的可能性和伤害的严重程度进行综合分析，评估风险大小的方法。它是一种风险可视化的工具，主要用于风险评估领域。该方法也可应用于电子档案管理业务的风险分析中。其主要思想是通过定性分析和定量分析综合考虑风险影响和概率两方面的风险因素，并评估风险因素对研究的影响。风险矩阵法的风险评估值由风险影响等级和风险概率两个因素构成，风险影响等级可以人为设定，风险概率可由样本数据测算得出。与电子档案管理业务相结合，以电子档案管理的审批流程为例，我们可以通过分析这些管理流程，进行风险预测，从而对电子档案的可信度进行评估。

5 可行性分析

5.1 深度分析法

深度分析法的最大值设置与风险矩阵法的风险影响等级设置类似，可参考《信息系统安全等级保护要求》中的风险等级设定方法，选取每项的最大风险等级，取值在1~5

之间，但是深度分析法的评估值大部分来源于定性的判断，其数字是有参与评定的人员在某一业务环节多次操作后所得到的一个主观评判值，该值一般不好通过样本直接给出，所以该方法得出的结果与实际值具有一定的偏差，需要经过

多次分析，不断积累学习后方可无限趋近于准确值。

选择电子文件归档和移交两个节点为度量节点为例，选择两个节点上部分行为活动进行分析并给出主观评定指标，具体见表 1。

表 1 电子文件归档和移交的主观评定指标

全周期	度量节点 (记为 N_i)	行为分类 (记为 C_{ij})	评估值 (记为 h_{ijk})	最大值 (记为 H_{ijk})
	权重为 $W_i\%$			
电子档案 (记为 A)	归档 (可信根) 权重为 $W_1\%$, 假定 $W_1\%=60\%$	电子文件信息	1.8	2
			4.5	5
			3	4
			2.5	3
		归档人员、档案管理员行为	3	3
			3	3
			4	4
	移交 (可信节点) 权重为 $W_2\%$, 假定 $W_2\%=40\%$	移交信息包	3.5	4
			3	3
			3.9	4
			4.8	5
		移交接收人员	4.9	5
			1	2
			2.5	3
		3	3	

将表 1 中的评估值和最大值代入公式得：

$$Ac_{11} = \frac{e^{1.8} + e^{4.5} + e^3 + e^{2.5}}{e^2 + e^5 + e^4 + e^3} = 0.5567$$

$$Ac_{12} = \frac{e^3 + e^3 + e^4 + e^{3.5}}{e^3 + e^3 + e^4 + e^4} = 0.8565$$

$$Ac_{21} = \frac{e^3 + e^{3.9} + e^{4.8} + e^{4.9}}{e^3 + e^4 + e^5 + e^5} = 0.8756$$

$$Ac_{22} = \frac{e^1 + e^{2.5} + e^3}{e^2 + e^3 + e^2} = 0.7357$$

则：

$$A_{N_1} = \frac{1}{j} \sum_j M_{C_{ij}} \quad A_{C_{ij}} \in (0,1]$$

$$A_{N_1} = \frac{1}{2} (M_{C_{11}} + M_{C_{12}}) = 0.7067$$

$$A_{N_2} = \frac{1}{2} (M_{C_{21}} + M_{C_{22}}) = 0.8056$$

从上述得值中可以看出，如果整个流程中仅有归档和移交两个环节的话，那么该系统的可信度为 $A=0.7463$ ，即 74.63%。

5.2 风险矩阵法

在风险矩阵法可信评价模型中，可以通过实例进行检测，来验证其可行性和科学性。项目组选取了南通市档案馆 10000 条 1983 年的文书档案样本数据，对其接收环节^[4]中的格式转换和四性检查两项行为活动进行了样本测试，测试情况如表 2 所示。

表 2 文书档案样本统计 (一万条 1983 年档案)

操作	统计内容	数量	转换成 pdf	
			成功	未成功
格式转换	doc 文件	0	0	0
	docx 文件	0	0	0
	pdf	3	3	0
	tif	9997	9997	0
	jpg	0	0	0
	数量		符合	不符合
四性检查	真实性	10000	10000	0
	完整性	10000	8640	1360
	安全性	10000	10000	0
	可用性	10000	9980	20

在表 2 中，完整性检查主要检查了档号，归档年度、保管期限、标题、全宗号、立档单位、责任者、成文日期、发文字号等字段，其中档号不符合要求的 46 条记录，成文日期不符合要求的 1011 条，归档年度不符合要求的 303 条。可用性方面损坏记录为 20 条。

对各节点的风险等级评定要根据各单位实际情况自行设定，项目设定的风险等级为 1~5 级，参考《信息系统安全等级保护要求》中的风险等级设定方法，对风险概率的取值需要建立一定的样本，这样才能对风险概率有一个科学的取值。本样例中设定其风险影响等级为 5，根据前面推导出的数学评价模型公式：

$$A = \frac{n}{\prod_{i=1}^n} \cos \frac{\pi}{4} R_i P_i$$

可推算出在这两个环节的可信度为0.8567,如表3所示。

将整个接收流程纳入可信评价的范围,可得出该档案管理接收环节其可信度见表4。

为了更明确重要节点对整体流程的可信度影响,以格式转换的风险概率变化来测算整个流程的可信度,如表5所示。

表5中,格式转换的风险等级为5,风险概率分别取了0.138、0.1、0.5、0.7四个值,代表其格式转换时发生的错误概率分别为13.8%、1%、50%、70%,那么它对整体业务流程的可信度影响是非常明显的,分别对应的值是0.9689、0.9749、-0.3734、-0.9014,从数据可以看出,当该风险概率达到50%时,得到的可信度为负数,说明已经完全不可信。

表3 两个环节的可信度

节点编号	业务事项	风险等级 Ri	风险概率 Pi	$\pi/4$	可信度
DY5	格式转换	5	0.0000	0.7854	1.0000
DY6	接收检测(四性)	5	0.1380	0.7854	0.8567
流程中节点的可信度					0.8567

表4 档案管理接收环节可信度

流程名称	序号	节点编号	业务事项	风险等级 Ri	风险概率 Pi	$\pi/4$	可信度
电子档案接收 业务流程	1	DY1	下发电子档案移交通知	1	0.0100	0.7854	1.0000
	2	DY2	提前移交申请	1	0.0100	0.7854	1.0000
	3	DY3	接收审核	4	0.0200	0.7854	0.9980
	4	DY4	移交数据预接收至临时库	3	0.0100	0.7854	0.9997
	5	DY5	格式转换	5	0.1380	0.7854	0.8567
	6	DY6	接收检测(四性)	3	0.0800	0.7854	0.9823
	7	DY7	部门负责人接收审核	2	0.0600	0.7854	0.9956
	8	DY8	办理外部交接文据	1	0.0100	0.7854	1.0000
整个流程的可信度							0.8358

表5 整个流程的可信度

流程名称	序号	业务事项	风险等级 Ri	风险概率 Pi	$\pi/4$	可信度
电子档案接收 业务流程	1	下发电子档案移交通知	1	0.0100	0.7854	1.0000
	2	提前移交申请	1	0.0100	0.7854	1.0000
	3	接收审核	4	0.0200	0.7854	0.9980
	4	移交数据预接收至临时库	3	0.0100	0.7854	0.9997
	5	格式转换	5	0.1380	0.7854	0.8567
				0.0100	0.7854	0.9992
				0.5000	0.7854	-0.3827
				0.7000	0.7854	-0.9239
	6	接收检测(四性)	3	0.0800	0.7854	0.9823
	7	部门负责人接收审核	2	0.0600	0.7854	0.9956
	8	办理外部交接文据	1	0.0100	0.7854	1.0000
	整个流程的可信度					
整个流程的可信度						0.9749
整个流程的可信度						-0.3734
整个流程的可信度						-0.9014

6 结语

论文通过采用深度分析法和风险矩阵法,构建了一个用于电子档案全生命周期四性检测的可信度测算模型。该模型利用数学方法和定量指标分析,能够评估电子档案的可信度,不仅适用于单个指标或业务环节的评估,也可用于电子档案全生命周期的可信度评估。此外,模型通过可信度测算,识别电子档案全生命周期管理和系统建设中的问题,并提出

相应的解决策略。

参考文献

- [1] 国家档案局.DA/T 70—2018 文书类电子档案检测一般要求[S].
- [2] DA/T 1—2000 档案工作基本术语[S].
- [3] 国家档案局.GB/T 18894—2016 电子文件归档与电子档案管理规范[S].
- [4] 国家档案局.DA/T 93—2022 电子档案移交接收操作规程[S].