doi: 10.7690/bgzdh.2024.07.006

基于模糊聚类算法的电子档案分类管理系统

郑黎明
（保定市人才交流服务中心，河北 保定 071000）

摘要：为增强电子档案资料管理水平，设计一种基于单标签射频识别（radio frequency identification，RFID）的电子档案分类管理系统。使用 RFID 搭建系统框架，通过中心数据库、终端管理器模块实现 RFID 数据存储与传输，运用文件管理、档案管理、开发利用及系统维护 4 个模块完成电子档案分类管理日常运维；引入模糊聚类算法提取电子档案数据信息熵，使用关联规则实现数据融合与自主调度，特征分解数据运行状态信息，并通过神经网络组建分类器对电子档案分类。实验结果证明：该系统运行时能实现高负载均衡，且 CPU 利用率低，在分类管理方面拥有准确率高、响应速率快等优势。

关键词：单标签 RFID；电子档案；分类管理；系统设计；模糊聚类
中图分类号：TP315　　**文献标志码：**A

Electronic Archives Classification Management System Based on Fuzzy Clustering Algorithm

Zheng Liming
(*Baoding Talent Exchange Service Center*, *Baoding* 071000, *China*)

Abstract: In order to enhance the management level of electronic archives, an electronic archives classification management system based on single tag radio frequency identification (RFID) is designed. Use RFID to build the system framework, realize the storage and transmission of RFID data through the central database and terminal manager module, and use four modules, including file management, file management, development and utilization, and system maintenance, to complete the daily operation and maintenance of electronic file classification management; The fuzzy clustering algorithm is introduced to extract the information entropy of electronic archives data, the association rules are used to achieve data fusion and autonomous scheduling, the feature decomposition data operation state information, and the neural network classifier is established to classify electronic archives. The experimental results show that the system achieves high load balancing, low CPU utilization, and has the advantages of high accuracy and fast response rate in classification management.

Keywords: single tag RFID; electronic archives; classification management; system design; fuzzy clustering

0 引言

全球化与信息化时代的到来，给电子档案管理带来新的发展与挑战，怎样提升电子档案管理服务效率[1]、更加贴合用户切实需求，是目前档案数字化建设领域关注的话题。档案管理过程中，按需分类操作十分重要，可将电子档案依照时间、内容、部门和载体[2]进行划分，不但方便查阅，而且能及时发现档案数据的不足，实现电子档案科学化管理目标。

关于数据分类问题，霍光煜等[3]使用隐含狄利克雷分布（Latent Dirichlet allocation，LDA）模型采集文档特征，通过 *K*-means 算法聚类主题特征，使用深度学习模型即可完成档案数据划分任务。朱赛赛等[4]提出一种融合全局和局部标记相关性的学习算法，运用余弦相似性明确不同标记的正负相关性，达到数据分类目的。

上述方法均存在计算量过高的问题，分类耗时长导致用户体验感较差。笔者依照企业单位实际需求，设计一种基于单标签射频识别（RFID）的电子档案分类管理系统。分析 RFID 技术操作过程，利用 RFID 数据管理子系统与档案信息管理子系统组建系统全局架构，使用模糊聚类法对电子档案进行分类，以提高分类精度。

1 基于单标签 RFID 的电子档案分类管理系统

RFID 技术的操作原理：当 RFID 标签卡处于读写器的射频范围内，并在该范围内输出射频信号到标签内，通过天线获取感应电流[5]，芯片内保存的产品信息通过内置天线自动发射，读写器把信息传

收稿日期：2024-03-25；修回日期：2024-04-28
第一作者：郑黎明（1984—），女，河北人。

送到计算机进行数据分析。

RFID 技术依照不同应用途径涵盖有源标签、无源标签及半无源标签。其无源标签虽然作用距离很短，通常为几十厘米至几十米区间范围[6]，但使用寿命长且应用范围较广，质量轻、价格低廉；因此，笔者将无源标签作为标记方式。

1.1 系统框架构建

基于单标签 RFID 的电子档案分类管理系统由 RFID 数据管理子系统与档案信息管理子系统共同构成[7]，全局架构如图 1 所示。

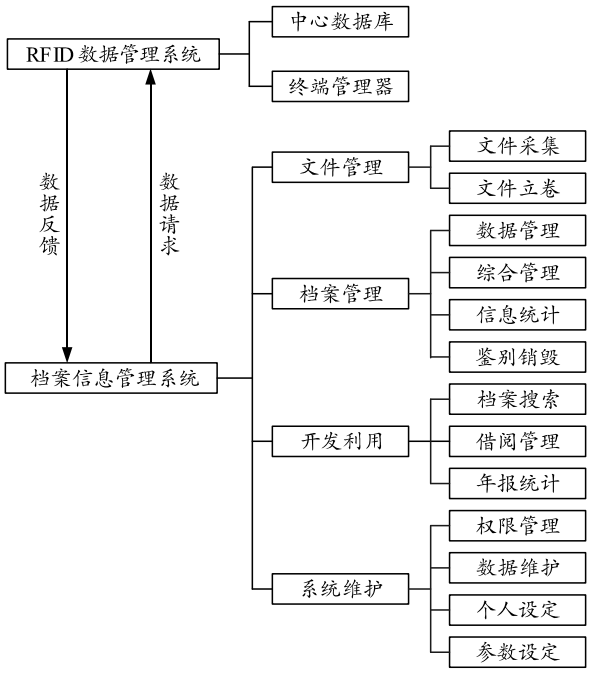


图 1 单标签 RFID 下电子档案分类管理系统架构

RFID 数据管理子系统包含中心数据库与终端管理器 2 个模块。其中，终端管理器模块通过读写器和手持式阅读器组成。读写器拥有信息导入功能[8]，手持阅读器是用户进行现场数据收集与通信的硬件，用户把收集的信息传递到中心数据库。

档案信息管理子系统拥有文件管理、档案管理、开发利用和系统维护 4 个模块，不同模块具备不同应用性能，为充分了解系统的应用模式，详细分析各模块功能。

1) 文件管理模块。

文件管理模块涵盖文件采集与文件立卷 2 个功能，点击文件管理进入文件管理界面，用户在此界面可完成对某个文件档案的收集规整工作[9]，点击文件采集按钮后，用户即可明确系统所属各部门供应的文件档案，并查看供应文件的单位名称；文件立卷页面中，可按照电子档案不同属性进行文件分类，实现电子档案归类任务。

2) 档案管理模块。

档案管理模块涵盖数据管理、综合管理、信息统计与鉴别销毁 4 个功能。进入档案管理页面后，用户在此页面可挑选已完成分类的文件档案，方便进行档案目录号、档案馆代号的查询[10]。综合管理功能中，可对电子档案的储存库房、储存设备等进行全方位管理设置。信息统计可将电子档案以统计图的形式呈现给用户，便于用户实时了解电子档案变化情况。

鉴别销毁功能中，对没有存档价值的档案要采取集中处理[11]，降低资源占有率。档案保存入档时，录入 RFID 标签并保存在中心数据库中。若超出保存时间，鉴别销毁功能会进行失效提醒，让用户进行销毁或继续保存。用户也能发出无线指令，明确 RFID 标签中的档案保管时间范围，完成对档案有效期的恰当管理。

3) 开发利用模块。

开发利用模块包含档案搜索、借阅管理、年报统计 3 个子功能。使用开发利用功能后，进入开发利用页面，用户在此页面完成档案开发。进入档案搜索界面时，输入想查找的档案数据，点击搜索完成档案检索。借阅管理页面中首先要进行用户登记，档案被借阅后标为已借出，归还时撤销借阅标签[12]，权限高的用户还能设置借阅时间和数量。年报统计以年为单位进行档案分析，通过报告的形式展现给用户。

4) 系统维护模块。

系统维护模块内含权限管理、数据维护、个人设定与参数设定 4 个功能。点击系统维护板块后，用户在此页面可管理全部调度流程。进入权限管理后，用户被赋予较高的权限[13]，查找权限高的档案或实施其他高权限操作。数据维护页面中，能完成电子档案系统日常运维修复。个人设定功能是完成人性化设置，譬如更改密码、用户名修改等。参数设定可以完成日期、卷号等档案编码设置，规范管理海量电子档案。

1.2 模糊聚类下电子档案分类算法

为保证电子档案分类系统运行可靠性，在系统中引入基于模糊聚类的电子档案分类算法，令其分类精度满足日常应用需求。

电子档案数据采集中，要明确电子档案数据的特征分布情况[14]，特征分布计算过程为：

$$A_{\infty}(b_{\mathrm{Th}})=1/(1+\exp((b_{\mathrm{Th}}+n)/n))。\tag{1}$$

式中：b_{Th}为电子档案数据模糊度；n 为电子档案总数；$A_{\infty}(\bullet)$为特征分布函数。

提取电子档案数据信息熵，使用关联规则挖掘策略进行电子档案数据融合与自主调度，使用运行状态监测法创建电子档案数据统计模型：

$$c(t)=\boldsymbol{D}_n\left[f(X,t)+e(t)+F\right]。\tag{2}$$

式中：$\boldsymbol{D}_n$为特征矢量；$f(X,t)$为档案分类监测点的模糊度；$e(t)$为分类函数；F 为特征统计值总和。

接下来使用空间分布融合策略实施电子档案数据高分辨率重组，重组后的数据主成分特征量为：

$$g(t)=\sum_{k=0}^{\infty}h_k(t)\cos\left[2\pi k f_m+l_k(t)+\theta\right];\tag{3}$$

$$h_k(t)=o\sum_{k=0}^{\infty}\boldsymbol{E}_{kn}\cos\left[2\pi k f_p+\varphi\right];\tag{4}$$

$$l_k(t)=\sum_{k=0}^{\infty}\boldsymbol{G}_{ki}\sin\left(2\pi i f_p+\varphi\right)。\tag{5}$$

式中：$\boldsymbol{E}_{kn}$、$\boldsymbol{G}_{ki}$均为电子档案数据的模糊特征矢量；θ 为电子档案数据的谱分解指数；f_m 为电子档案数据采样率；f_p为电子档案数据状态点频率；$h_k(t)$、$l_k(t)$依次为电子档案数据的采样速率与采样数量；φ 为数据显著性特征。

采样分布范围内，将电子档案数据状态监测公式记作：

$$y(t)=\left[1+E\cos\left(2\pi f_p+M\right)\right]\bullet\cos\left[2\pi f_p+G\sin\left[2\pi f_m+\varphi\right]+M\right]。\tag{6}$$

式中 M 为电子档案数据谱峰值。

综上所述，将电子档案数据模糊特征向量记作：

$$\hat{\boldsymbol{H}}_j=\begin{cases}Q_j\left(1-\rho^2/\mathrm{Sum}_j^2\right), & \mathrm{Sum}_j^2\geqslant\rho^2\\ 0, & \mathrm{Sum}_j^2<\rho^2\end{cases}。\tag{7}$$

式中：Q_j为模糊指数；Sum_j^2为特征评估阈值；ρ^2为电子档案分类临界值。

特征分解电子档案数据运行状态信息，引入统计特征聚类算法[15]，获得电子档案分类矩阵 $\boldsymbol{Y}$ 的奇异值分解结果，如式(8)所示。

$$\boldsymbol{Y}=\boldsymbol{W}PV^2。\tag{8}$$

式中：$\boldsymbol{W}$ 为特征分布正交矩阵；V 为聚类中心；P 为数据子带分布维数。

利用神经网络法创建电子档案分类器，如图 2 所示。

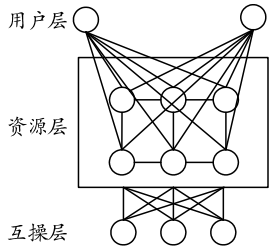


图 2　神经网络分类器

上图中，用户层内包含不同领域的数据资源与网络检索工具，实现档案分类资源聚合，给用户提供性能优秀的导航工具，提升档案搜索性能；资源层使用试点运行方式构建数据整合准则，涵盖集合级描述、对象级描述和子资源聚合描述，利用此类描述完成分布式信息搜索与资源定位；互操作层中，异构数据之间的互操作与资源对象描述不同，拥有分布式特征，具备较多层次，描述的目标处于不同信息系统。此类描述目标的格式、数据架构等均为异构的，完成异构资源目标无缝衔接，达到异构信息系统之间的互操作，分类过程更为流畅，极大提高档案分类效率。

设置电子档案数据特征值为 $\mu_1,\mu_2,\ldots,\mu_i$，将其引入神经网络分类器中，利用空间特征点分布误差衡量其数据属性特点[16]，通过式(9)输出档案分类结果，实现科学化电子档案管理目标。

$$\mathrm{SURE}(\omega_m,\boldsymbol{L})=R_m+\sum_n\left\|r_n(f_s)\right\|_2^2。\tag{9}$$

式中：ω_m为电子档案数据分布特征集合；$\boldsymbol{L}$ 为电子档案数据特征向量；R_m 为档案数据自相关特征；f_s为数据融合特征集合；r_n为关联规则系数。

2　实验分析与结果

2.1　实验设置

用实验分析来证明所建系统的有效性，并将其与 LDA 模型法[3]、标记相关性法[4]进行对比，对比指标为分类精度、分类效率、不同分类管理状态下计算机的 CPU 利用率与负载均衡情况，让实验结果更具客观性。实验平台为 Matlab 7.0，实验数据为某物流企业 2021 年度的电子图像数据文件，包含 15 000 条档案信息，每条档案的规格各不相等。

衡量方法分类精度时，采用分全率、分准率和宏观 F 值进行性能评价。分全率表示数据集内全部档案被正确分类个数所占的比例；分准率表示分类的档案和全部档案的百分比；宏观 F 值是对分全率与分准率进行综合估算的判断指标。分别把 3 个评价指标的计算过程描述为：

$$RI=TM/FB; \tag{10}$$

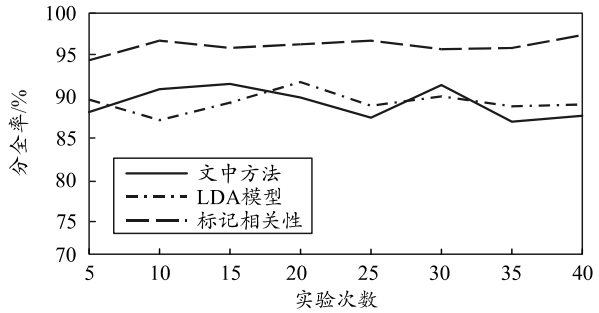
$$PI=TM/FC; \tag{11}$$

$$F=RI\times PI\times 2/(RI\div PI)。 \tag{12}$$

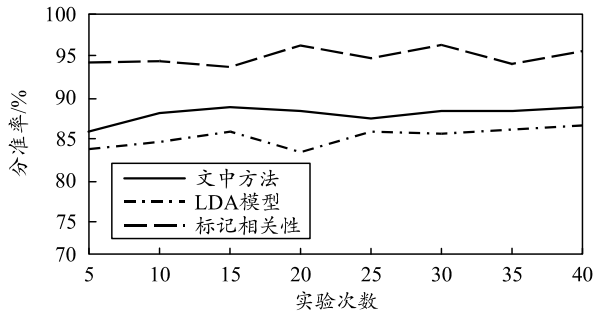
式中：TM 为被准确分类至自身属性相关的档案数量；FB 为该属性电子档案真实数量；FC 为数据集被划分到相同类型的档案数量。

2.2　实验结果

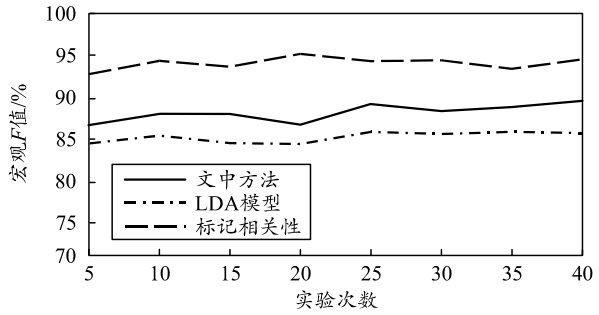
3 种方法电子档案分类精度实验结果如图 3。



(a) 分全率性能对比



(b) 分准率性能对比



(c) 宏观 F 值对比

图 3　电子档案分类精度指标对比

观察上图可知，在不同指标对比分析中，文中方法的分全率和分准率均不同程度地大于 2 个文献方法，宏观 F 值也随之提高，分类精度更胜一筹。这是因为文中方法采用单标签 RFID 技术，在整理电子档案时能快速锁定档案的关键信息，显著改善电子档案整合归类准确性。

信息化时代，电子档案分类耗时的多少直接决定了其管理模式是否能投入使用，这也是档案管理中最为核心的评估指标。把 15 000 个档案分为 5 组，每组档案个数为 3 000 个，取各组分类时间均值为对比目标，探讨不同方法下电子档案分类效率优劣，结果如图 4 所示。由此看出，伴随实验数量的递增，文中方法分类耗时没有显著变化，所耗时长一直处于最短，LDA 模型法次之，标记相关性法分类时间最长，这也进一步验证了所建系统的适用性。

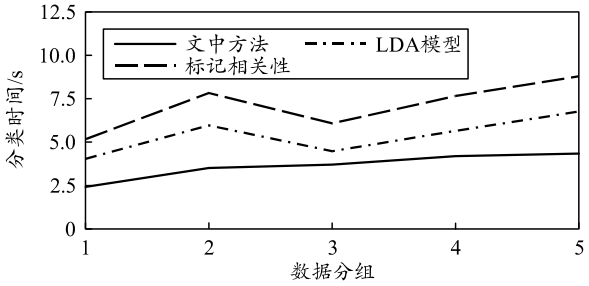


图 4　电子档案数据分类时间对比

设置用户承载量为 600 人，对比 3 种方法电子档案分类时计算机的 CPU 利用率情况，CPU 利用率越低，证明计算机运行越流畅，方法的实用性越强，实验结果如图 5 所示。

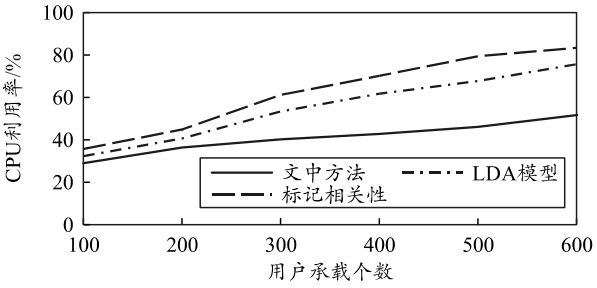


图 5　电子档案分类 CPU 利用率对比

从上图可知，在用户数量较少时，3 种方法的 CPU 利用率无显著差异，但用户数量大于 300 人后出现较大变化，文中方法 CPU 利用率要低于 2 个文献方法，证明其在多用户并发使用下不会延误计算机的正常运行。

通过负载均衡来表明 3 种方法下计算机数据处理能力，假设电子档案分类管理时具备 7 个服务器，探究 3 种方法下档案分类过程中服务器的使用情况，如图 6 所示。

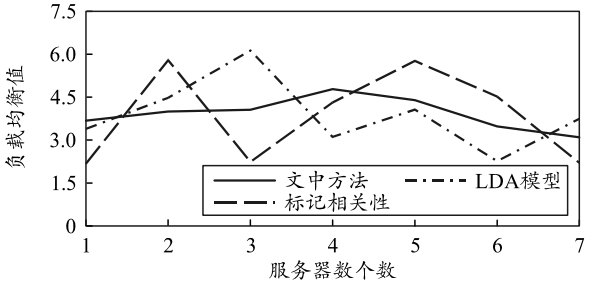


图 6　电子档案分类负载均衡对比

（下转第 51 页）