基于结构化和文本数据的辅助开放鉴定模型应用

（2023 年度国家档案局优秀科技成果二等奖）

一、成果简介：

课题以综合档案馆馆藏新中国成立后纸质文书档案（以下简称文书档案）为主要研究对象，采用文献研究、德尔菲法、实验和实证等方法，从分析档案信息特征入手，率先提出了从 16 个维度辅助开放鉴定的方法，综合运用数据挖掘等人工智能技术、关键词匹配技术、档案知识服务和专家经验，构建了由递进式辅助开放鉴定双模块、机器学习与深度学习等相关算法模型、深度学习神经网络模型，以及档案知识库组成的辅助档案开放鉴定模型（以下简称辅助开放鉴定模型），具有创新性，且辅助开放鉴定模型具有弹性和可扩展性，形成并实现了智能辅助档案开放审核一整套解决方案。课题研究期间，在江西省档案馆文书档案开放审核工作中开展了辅助开放鉴定模型实证研究，证明了它的先进性、可用性。研究成果主要有 4 个方面：

1.**构建递进式辅助开放鉴定双模块及其预测流程。**经过实验分析，课题筛选了特定关键词等 5 类档案信息特征，并将其转化为 16 个分析维度，分别组合构成两个辅助开放鉴定模块，并可根据实际需要增加新的维度或删除原有维度。(1)单维直接鉴定模块。该模块由特定开放词等 7 个维度组成，基于相应特定关键词库等直接预测档案开放与否，是对以敏感词辅助“划控”为代表的关键词匹配方法的深化与拓展。(2)多维加权组合鉴定模块。由开放文本相似度、控制文本分类和情感分析等 9 个维度构成，运用经过测试、调参和训练的 LSI、TextCNN、BERT+Dense 等算法模型分别计算维度特征值；经反复实验、测试，创建含 4 个隐藏

层的深度学习网络感知器二分类模型，用于计算 9 个维度动态权重系数并生成预测结果，同时采用置信度作为预测意见强度参数。(3)辅助开放鉴定模型预测流程。预测流程首先进入单维直接鉴定模块，其任一维度输出预测结果，则预测流程结束，否则进入多维加权组合鉴定模块。

2.**构建辅助开放鉴定模型档案知识库。**(1)构建辅助开放鉴定模型底层档案知识库。档案知识库由各种高质量档案词库、档案语料库、档案查重库等组成，它服务于上层的递进式辅助鉴定双模块，支撑辅助开放鉴定模型相关维度运算和相关模型的学习、训练等。(2)创建档案开放词库和敏感词库。创造性运用中文分词、散点图等技术创建了档案开放词库与敏感词库，设计词条元数据，研究形成新词确认规则。(3)实施文书档案数据化。课题采用百度飞桨平台对纸质档案数字化副本进行 OCR 识别，将原文数据化并存入中间数据库供数据挖掘分析使用。

3.**构建并实用了跨网络、跨平台、跨系统的体系化辅助开放鉴定模型系统功能。**（1）辅助开放鉴定模型与应用功能实现。课题依托江西档案云中心的计算和存储资源、档案大数据分析应用系统、省档案馆数字档案集成管理系统、OCR 等，以基于 MPLSVPN、国密信道等技术部署的档案纵向业务网及其网络安全防范体系为支撑，构建、开发了辅助开放鉴定模型、待预测档案数据推送与采集、OCR 识别、中文分词、预测成果反馈与实际应用等一系列系统和功能，完成了档案大数据分析应用系统、中间接口、数字档案集成管理系统等第三方应用系统对接，打通了开放审核业务流与档案数据流通道，为课题研究与实证提供了技术支撑和保障。(2)NLP 工具造型与二次开发。课题测试选用了 PaddleOCR 和 LAC 等两款必备工具，均为开源产品，根据课题研究和后期实用需

要分别做了应用层二次开发和微调。（3）建立可实用的人工智能相关开源技术“工具箱”。经测试选型，有效应用了包括 scikit-learn 机器学习库、Gensim、机器学习框架 TensorFlow、BERT 预训练语言表征模型，以及数据分析工具 Pandas 和数据可视化工具 Matplotlib、Tensorboard 等在内的相关开源技术工具，与 PaddleOCR、LAC（微调）一并构成课题研究和成果应用的开源技术“工具箱”，有效降低研究和实用成本。

4.形成 13 万字学术研究报告和论文等。（1）在课题研究期间及完成后，项目组双方联合撰写 6 万字研究报告，主要包括绪论、研究现状、辅助开放鉴定模型构建需求、辅助开放鉴定模型构建与管理、辅助开放鉴定模型的实现等 5 章。（2）作为规范课题实验的重要措施，课题组随各种实验及时建立 6 万字主要实验记录，如实记录了中文分词与 OCR 工具、算法模型、技术路线选型以及维度筛选等实验过程和结果，为课题研究提供了重要依据和参考。（3）在 2021-2023 年《中国档案》杂志先后发表《数据挖掘技术在档案开放鉴定领域应用初探》等 3 篇论文，在 2022 年 8 月 1 日的《中国档案报》第 3 版之《好风凭借力 扬帆正当时——江西省档案馆推进馆藏文书档案开放审核工作纪实》中介绍课题研究成果应用情况。

二、成果详细内容：

1.辅助开放鉴定模型构建思路

档案开放鉴定是指依据相关法律法规，通过规范的工作程序，对封闭期已满的档案进行鉴别，确定档案可否对外开放以及如何开放的过程。因而，相关法规依据、识别档案开放和敏感信息特征是开放鉴定的两个关键点，前者是辅助开放鉴定模型应遵循并实现的，可称之为开放

鉴定规则库，后者是该模型应能识别的。经过多回合分析、实验，形成了辅助开放鉴定模型构建思路，即要在相关法规框架下，结合专家经验和数据挖掘等技术原理，形成开放鉴定规则特征、档案开放和敏感信息特征，并为其选择、训练相应的数据分析算法，实现二者的算法转换，据此设置模型维度、运算规则及其参数等。

2.确定辅助开放鉴定模型维度设置

2.1 开放鉴定规则特征

开放鉴定规则特征主要有关于涉密档案管理、“划控二十条”、应主动公开的政府信息类别等法规条款，以及综合档案馆制定的划控实施细则等，这些特征一并构成了开放鉴定规则库。根据实验，25 年档案封闭期、经济等类档案可提前开放等两条规则特征暂未找到可行的算法转换路线；关于涉及国家安全等档案可以延期开放的规定，只要辅助开放鉴定模型能够识别并为控制使用即可达到延期开放目的，为此，这 3 项规则特征未纳入开放鉴定规则库。

开放鉴定规则特征的算法转换原理是，用标注有“划控二十条”、政府信息主动公开类别等的档案数据语料，训练文本分类算法模型，选用已著录了开放或控制标识的档案数据二分类语料训练文本聚类算法，再分别执行开放鉴定对象的分类、聚类预测任务，给出预测结果及开放或控制程度的特征值。根据实验和应用情况，前者准确率更高，但激活率较低。结合算法转换和档案数据语料供给等实际，实提炼形成了密级、控制文本分类、控制文本相似度、开放文本相似度、公开信息文本分类等 5 个开放鉴定规则类维度。

2.2 档案信息特征

档案信息特征分为开放和敏感信息特征，是指开放或控制使用档案特有的内容和形式特征。经分析和实验，本课题形成了公开属性、互联网属性、开放词、特定开放词、特定开放文种等 5 种档案开放信息特征，敏感词、特定敏感词、特定控制文种、特定责任者、全宗敏感性等 5 种档案敏感信息特征，以及同时具有开放和敏感属性的档案信息情感特征，并设置了相应的辅助开放鉴定模型维度，其中，开放词、敏感词、档案信息情感特征等三个特征对应的维度是开放词数、敏感词数和情感分析，其他维度名称与对应的特征名相同。

2.3 辅助开放鉴定模型维度

本课题为辅助开放鉴定模型设置了 16 个维度，其中，7 个维度专用于开放方向，8 个维度专用于控制方向，情感分析维度则兼而用之。辅助开放鉴定模型的维度，是从不同角度直接或间接揭示档案信息开放或敏感程度的特征。下面从维度名称、含义、原理、值域等 4 个属性描述 16 个维度：

（1）公开属性，依据《信息公开条例》为公文设定的公开级别；由辅助开放鉴定模型分析、匹配开放鉴定对象的公开属性元数据、原文文本中的公开属性值；值域：主动公开，依申请公开，不予公开；

（2）开放词数，档案题名中含有的开放词数量；基于开放词数、开放词阈值计算形成、用于反映档案开放程度的一个数值;值域：无；实验表明，针对档案原文文本计算开放词数，受到文本长短、OCR 识别质量等因素的干扰；本维度计算以档案开放词库为基础，基于开放词频在不同档案数据语料中的分布情况，选取具有代表意义的中位数作为开放词数的阈值，在档案题名中抽取的开放词数量除以相应的阈值所获得

数值即为该维度的运算结果；

（3）特定开放词，对开放鉴定具有特定作用的开放词；档案题名中只要出现一个特定开放词即能够直接鉴定该件档案为开放；值域为特定开放词库中的所有开放词，例如：职务任免.....通知、表彰.....决定等；

（4）特定开放文种，对开放鉴定具有特定作用的公文文种；只要在档案题名中成功匹配特定开放文种，即可直接划定该档案为开放；值域为特定开放文种库中的所有文种，如公报、办法、条例、令等；

（5）开放文本相似度，开放鉴定对象与开放档案的相似程度；由聚类分析算法模型对开放鉴定对象做分析、运算，生成一个 0～1 之间的数值；值域：无;

（6）公开信息文本分类,开放鉴定对象与主动公开的某类政府公开信息的相似程度；由分类算法模型对开放鉴定对象进行分析、计算，生成一个 0～1 之间的数值；值域：无；

（7）互联网文本相似度，开放鉴定对象与网页、社交媒体等类发布在互联网的政务信息的相似程度；由聚类分析算法模型对开放鉴定对象做分析、计算，生成一个 0～1 之间的数值；值域：无；

（8）密级，档案保密程度的等级；由辅助开放鉴定模型匹配开放鉴定对象的密级元数据项，分析、比较开放鉴定对象或关联档案有无涉密信息；值域：密秘，机密，绝密；

（9）敏感词数，档案题名中含有的敏感词数量；基于敏感词阈值计算形成的开放程度、用于反映档案敏感程度的一个数值,阈值计算原理与开放词数相同；值域：无;

（10）特定敏感词,对开放鉴定具有特定作用的敏感词；档案题名中

只要出现一个特定敏感词即能够直接鉴定该件档案为控制；值域为特定敏感词库中的所有敏感词，例如干部处理、劳资纠纷、开除学籍等;

（11）特定控制文种，对开放鉴定具有特定作用的公文文种，只要在档案题名中成功匹配一种特定控制文种，即可直接划定该档案为控制；值域为特定控制文种库中的所有文种，如会议记录、介绍信存根等;

（12）特定责任者,非本地区相关单位制发的文件的责任者；运行代码脚本或特定责任者库，匹配责任者、题名中的特定字、词，例如，对省级综合档案馆馆藏档案，需匹配“中央”、“国务院”、“部”等字、词；值域为特定敏感词库中的相关敏感词;

（13）控制文本分类,开放鉴定对象与“划控二十条”任一类目中档案的相似程度；由经过训练的分类算法模型对开放鉴定对象进行分析、计算，生成一个 0～1 之间的数值；值域：无;

（14）控制文本相似度,开放鉴定对象与控制使用档案的相似程度；由聚类分析算法模型对开放鉴定对象做分析、计算，生成一个 0～1 之间的数值；值域：无;

（15）全宗敏感性,全宗档案整体敏感程度；全宗内控制使用档案在封闭期已满档案中所占比例达到设定标准以上的为敏感全宗，需要定期或不定期地计算、更新控制使用档案占比和敏感全宗名单；值域：1，0，分别用 1、0 代表敏感全宗和处于设定标准以下的全宗；

因立档单位职能等因素，许多全宗受控文书档案占比较大，有的全宗已完成开放鉴定文书档案中控制使用的占比在 60%以上，部分小全宗甚至达到 100%，具有敏感信息特征；经过实验分析，全宗敏感性特征对辅助开放鉴定模型的分析具有一定的影响；

（16）情感分析，档案数据中带有的正面或所谓负面情感色彩、情感倾向性等，由由语义分析算法模型分析、计算，生成一个 0～1 之间的数值；值域：无。

3.确定辅助开放鉴定模型关键技术选型

3.1 OCR 识别技术

课题组对 Tesseract 和 PaddleOCR 做了比较研究，多轮测试后选定深度学习架构的 PaddleOCR 为 OCR 识别工具，并根据档案数据化和应用系统对接需要做了二次开发,使其支持 PDF 格式数字化副本的自动拆分、自动批量 OCR 识别、并将 OCR 结果存入文本文件、数据库 blob 字段供系统调用。根据江西省档案馆持续使用情况来看，完成识别 70 余万件 PDF 格式数字副本,成功率为 93.7%，采用 16 线程同时工作时，平均每画幅耗时约为 3.36 秒，同时该工具的预测、识别质量与效率还将随使用时间的增加而持续提高。

3.2 中文分词技术

以自动分词为主、人工补充为辅的方式创建档案词库，亦属本课题研究内容。选用开源“cws_evaluation：中文分词器分词效果评估”数据集，对 Jieba、HanLP、LTP、LAC 等 4 种开源分词工具进行测试，采用精准率和召回率的调和平均 F1 值为主要测评指标，分词速度为次要评判指标。对上述 4 种开源分词工具以及经微调、增量训练的 LAC 模型做实验，测试表明微调后的 LAC 模型分词效果相对更好，F1 值为 0.95，分词速度为 37.1/每字符毫秒。该工具可微调、可学习优化的扩展性、成长性更符合综合档案馆档案数据化的持续中文分词需求。

3.3 辅助开放鉴定模型维度算法选型

开放词数、敏感词数等 2 个维度使用 DFA 搜索算法，主要针对文本聚类、文本分类和情感分析等 3 类算法开展维度算法选型实验，并为算法选型和后续模型集成调试准备了 162.8 万件文书档案长文（数字原文）、短文（目录数据）语料，17.6 万件政府公开信息、网页和互联网开源情感分析语料等 3 类语料，其中包括 3.2 万件三元组语料和 5.27 万件 27 类控制适用规则的标注语料。

（1）文本聚类算法选型。采用 scikit-learn 机器学习库，重点对 DBSCAN、LSI 等两种浅层学习算法开展实验，期间使用 Gensim 获得文本隐层主题向量表达，测试语料包括含有开放或控制标识的文书档案文件级目录数据各 2000 条、网页等政务信息语料 2000 条（件）。根据实验，LSI 聚类算法对开放文本相似度、控制文本相似度和互联网文本相似度的测试结果远远优于 DBSCAN 算法，平均分分别为 0.978、0.978、0.883，耗时分别为 2.347、2.81、2.381 秒。

（2）文本分类算法选型。实验工具增选了机器学习框架 TensorFlow、BERT 预训练语言表征模型，以及数据分析工具 Pandas 和数据可视化工具 Matplotlib、Tensorboard。

对公开信息文本分类维度，测试朴素贝叶斯、TextCNN 等两个算法模型，选取政府公开信息分类语料 9.51 万件，其中，7.3 万件做训练语料，0.73 万件用于验证，1.46 万件用于测试。根据测试，TextCNN 算法更优，平均正确率 0.758，耗时 2.346 秒。

对情感分析维度采用语义分析技术路线，选型围绕 BERT+Dense、ALBERT+Dense 、 one hot+LSTM 、 embedding+LSTM 、 word embedding+LSTM 等 5 种算法开展实验，选用互联网开源情感分析语料

2.11 万件，其中正面 1.06 万件、负面 1.04 万件，实验分三步进行。首先，训练 5 种算法，按 9：1 比例将上述语料随机分为训练集和测试集，导入 5 种算法模型分别进行多轮训练、测试，直到运算数据走向趋于稳定。根据实验，最慢的超过了 25 轮测试，BERT+Dense 算法完成 4 轮即趋于稳定，训练集、测试集准确率就分别达到 99%、95%，损失函数降至最小值，选择 BERT+Dense 算法模型进行后续验证实验；第二步是使用 1000 件政府公开信息语料验证 BERT+Dense 算法情感拟合分布，句子分布散点图证实了政府公开信息语料的正面情感；第三步，随机抽取开放、控制档案文件级目录各 2000 条，进一步分析该算法情感拟合，实验结果是正面情感取向的开放档案目录占 64.36%,大于控制档案目录正面情感取向的 10.23%，具有合理性。

对控制文本分类维度，鉴于 TextCNN 算法模型在处理类似档案内容上下文关系方面存在不足，再次采用 BERT 算法开展实验，选取文书档案二分类、27 类控制适用规则标注等 2 批语料，分长文、短文作 4 次测试，实验结果是二分类长、短文语料测试数据相对更优，正确率分别为 85.1%、76.5%，证实了该算法的可用性。

3.4 构建形成人工智能相关开源技术“工具箱”。

课题研究过程中，除了相关算法选型、调参等研究工作外，还根据相关工具和算法选型、维度测试与分析、技术路线比较分析等需要，逐步应用了诸多人工智能相关开源技术工具，包括 NLP 工具，例如 PaddleOCR、LAC；机器学习库、框架和模型，例如 scikit-learn、Gensim、TensorFlow、BERT 预训练语言表征模型；数据分析和可视化工具，例如 Pandas、Matplotlib、Tensorboard 等。这些开源工具经过二次开发、

调参和优化后，汇聚形成了智能化辅助档案开放审核所需的人工智能相关开源技术“工具箱”。

4.辅助开放鉴定模型构建与管理

4.1 维度运算方法与流程

采用 16 个维度实现二分类预测的方法可分为直接预测和间接预测两种。直接预测是指，通过匹配、比较档案所含开放或敏感信息特征，进而给出开放或控制使用的辅助鉴定意见，主要应用在公开属性、特定开放词、特定开放文种、密级、特定敏感词、特定敏感文种、特定责任者、全宗敏感性等 8 个维度，该方法相对简单且预测准确性较高。间接预测是指，运用经过训练的维度算法模型分析、计算开放鉴定对象，分别给出其开放或敏感程度，综合相关维度计算结果后给出二分类预测结果，当直接预测没有结果时，则需采用该方法，主要用于其他 8 个维度。

维度运算全过程可分为两段。第一段称为单维直接鉴定，适用直接预测方法，有任一维度激活并给出预测意见，辅助开放鉴定流程结束，否则，进入下一阶段；第二段称为多维加权组合鉴定，适用于间接预测法，经过计算和分析得出各维度能够反映档案开放或敏感程度的一个数值，模型自动将各维度值及其权重系数代入数学公式进行计算，得出二分类预测结果。两个阶段以及相应的算法模型、分析过程构成了辅助开放鉴定模型的核心数学模型，称之为递进式辅助开放鉴定双模块，由于集成的每个或一类维度的含义、设置目的、实现原理有所不同，增强了综合运算所得预测结果的可信度。

4.2 构建辅助开放鉴定模型

递进式辅助开放鉴定双模块是以下层档案知识服务为支撑，通过维

度算法模型和机器学习工具、深度学习神经网络驱动，融合完成识别、分析和运算任务，按规则给出开放或控制的预测结果，这个过程使得档案知识库不断丰富，维度算法模型将档案知识转化、提炼并上升为智慧。由此，将辅助开放鉴定模型构建为由下至上的知识层和智慧层等两层结构，如图 1 所示。知识层负责为智慧层提供档案知识服务，由各类档案知识库构成。智慧层负责向第三方应用系统输出辅助开放鉴定意见，由递进式辅助开放鉴定双模块、维度算法模型组成，单维直接鉴定和多维加权组合鉴定构成递进式辅助开放鉴定全流程；维度算法模型负责完成多维加权组合鉴定的识别、运算和分析任务，通过持续学习、优化和提升，维度算法模型呈螺旋式成长态势，思维能力和智慧水平不断提高。

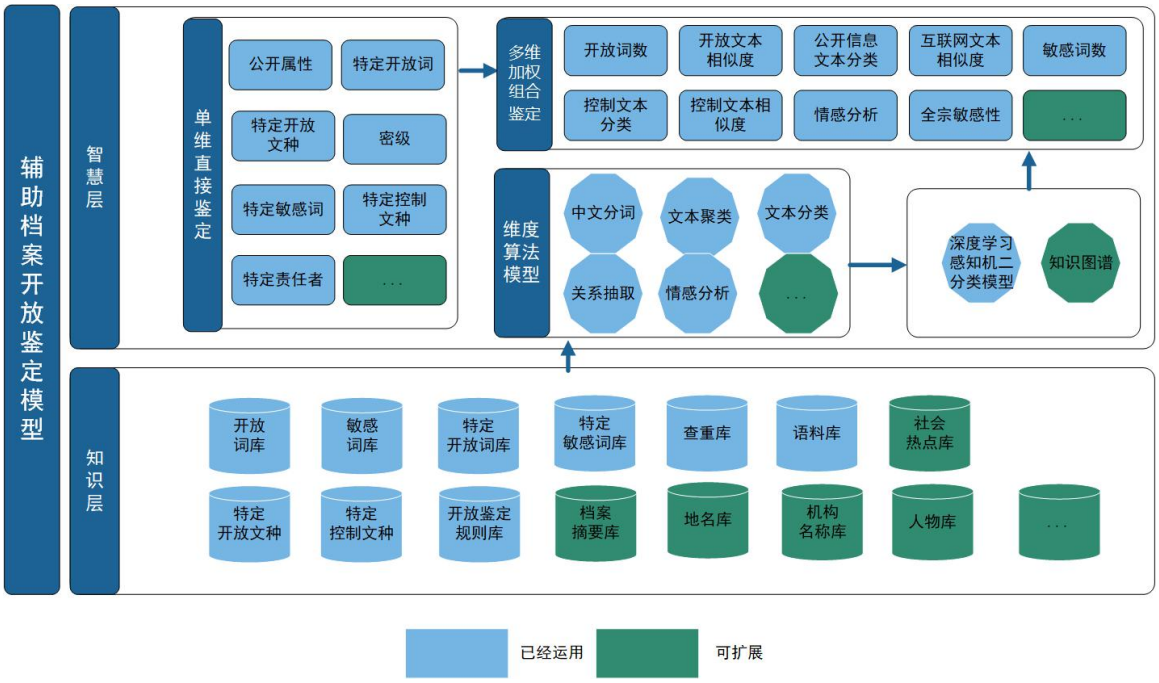

图 1 辅助开放鉴定模型示意图

在辅助开放鉴定模型中，两种特定词库、两种特定文种库分别服务于四个相应单维直接鉴定维度，开放词库、敏感词库分别为开放词数、

敏感词数等两个维度提供服务，开放鉴定规则库是控制文本分类、公开信息文本分类等维度的预测规则，语料库直接服务于档案词库创建任务和多维加权组合鉴定各个维度；档案摘要库由摘要算法生成，可传输给第三方应用系统供开放鉴定人员参考；查重库存储的是域内跨馆、跨全宗重复档案目录及其标记，有助于优化开放鉴定计划，有效减少开放鉴定工作量。

辅助开放鉴定模型具有良好的弹性、扩展性和成长性。第一，根据辅助开放鉴定对象的具体情况，可以从递进式辅助开放鉴定双模块中移出或增加若干维度，只要模块集成测试达到预期目标便可投入应用。例如，辅助民国档案开放鉴定任务，可从单维直接鉴定模块中移出公开属性维度，将公开信息文本分类、互联网文本相似度等两个维度从多维度加权组合鉴定模块中取出。第二，可以根据维度设置和维度算法训练的需要，在知识层注入新的知识库。例如，在多维加权组合鉴定模块中新增隐私信息相似度及相应算法模型，便可通过新增人脸特征库、人物库、机构名称库、地名库等知识库为个人（机构）隐私信息相似度算法模型服务。第三，随着辅助开放鉴定任务的持续进行，高质量档案数据语料不断输入，档案词库和语料库逐渐成长、充实，维度算法模型越来越聪明，使得辅助开放鉴定模型始终处于成长之中。

4.3 辅助开放鉴定模型集成测试与优化

（1） 档案词库的优化

创建词库初期，从开放档案获取的新词写入开放词库，从控制档案切分所得新词放入敏感词库，使两种词库迅速膨胀，词数很快超过 80 万个，敏感词库中的人名、地名、机构名等词以及大量两库重叠词对预

测产生明显噪声，例如，开放词数、敏感词数常被同一个词激活。为此，从两个方面优化新词入库规则。首先，明确词库定位，开放词仅出现在开放档案，敏感词仅出现在控制档案中。其次，研究新词入库规则。提取 4 种门类约 400 万件档案数据语料分词，分别从控制档案、开放档案中抽取词（含短语）共计 54.59 万个、163.49 万个，其中，重叠部分 51.95 万个，有 10.34 万个词的词频小于 2；创建词分布散点图，用 X 轴、Y 轴分别代表某词在开放档案或控制档案中出现的频率，如图 2 所示；将全部词面积计为 1，切去重叠词占比，得出划分开放词、敏感词和停用词的阈值为 0.08；取符合 x-y<=0.08 的词，即散点图对角线附近的词归集于停用词库中；符合 x>y 且 x>y+0.08 的词，即对角线下方的词划分为开放词；符合 x<y 且 x<y-0.08 的词，即对角线上方的词划分为敏感词；按照从严标准，在运用阈值基础上，再过滤词频小于 2、对辅助开放鉴定构成轻微影响的词。依照上述规则创建的档案词库迅速“瘦身”，噪声显著降低，截止 2021 年 12 月课题验收前，文书档案开放词库共 2.44 万个词，敏感词库共 9.3 万个词。

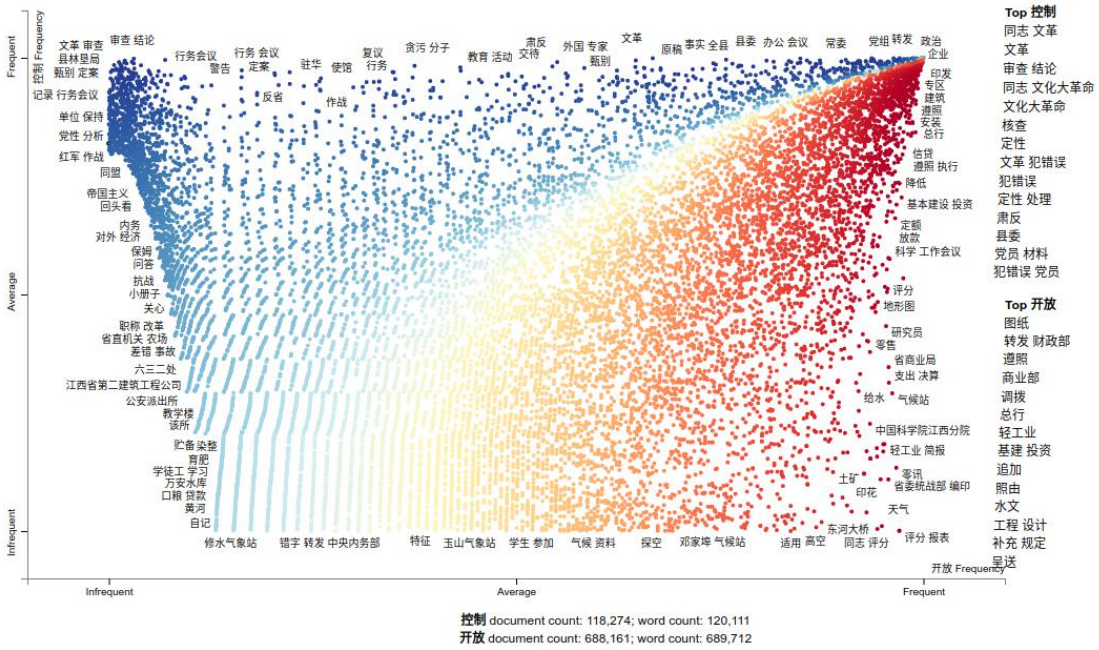

图 2 通过 Scattertext 分析档案文本的可视化结果

（2）单维直接鉴定模块测试与调优

首次对单维直接鉴定模块的 7 个维度测试结果很不理想，使用 3.76 万件文书档案语料开展实验，7 个维度共命中了 9077 件档案，与语料控制标识对比一致的只有 4350 件，总准确率为 47.9%。经过分析，主要有分词错误、特定敏感词设置不当、非本地区责任者识别错误等问题，例如，特定敏感词库中的“批复”一词误将 40 多件档案划为控制使用，说明单维直接鉴定意见准确与否，与知识库质量高低密切关联。经过多次优化，后期单维直接鉴定准确率大幅提升，提取文书档案 2.5618 万件、1.3427 万件等两批语料做实验，7 个维度分别命中 8067 件、2728 件，与高质量档案语料控制标识相比较，对比一致的分别为 7962 件、2496 件，准确率为 98.71%和 91.53%。

（3）多维加权组合鉴定模块测试与优化

针对多维加权组合鉴定模块，先后采用专家经验、广义线性回归和深度学习路线测试计算方法和权重配置，经过实验，深度学习方法相对更优。

a.专家经验法。在课题研究较长一段时间内，一直采用点数、引入双曲正切函数计算可信程度的方法，例如，控制点数=Σ维度*权重系数（控制方向），当控制点数>开放点数时，辅助鉴定为控制，当开放点数>控制点数时，辅助鉴定为开放；再采用公式可信程度=tanh|控制点数-开放点数|计算可信程度，用百分比表示，值域为 0—100%。虽然人工对各维度权重系数作了数次调整和实验，批次语料辅助鉴定意见准确率一直在 50%左右徘徊，课题组认为语料控制标识存在噪声、维度权重系数设置不合理等问题是造成准确率偏低的主要原因。为此，在全国档案专家库中邀请多个专业方向的 10 位专家，就维度权重配置开展了两轮德尔菲咨询。根据咨询成果，课题组做了 3 种权重对比实验，准确率最高只有 52.47%。

b.广义线性回归法。专家经验法遇到瓶颈后，课题组分析认为维度之间存在由权重系数串起的“线性关系”，选择广义线性回归法作进一步实验，该法是运用数理统计中的回归分析来确定两个或多个变量相互依赖的定量关系。选用 2.7 万件标注了 27 类控制适用规则的文书档案语料，借助历史试验数据和 rapidminder 工具对多维度样本进行分析，构建并运用 GLM 模型计算权重系数并给出预测结果。初次实验取得明显突破，3 种点数与权重配置实验准确率最高为 84.65%。但是，后续多次实验结果出现反转，准确率都在 55%左右。再次提取 3.7 万件文书档案二分类

语料，使用 t-SNE 算法分析,通过非线性降维处理后的可视化图像（图 3）显示，开放或控制档案以多维、非线性形态存在，若使用广义线性回归法执行档案二分类预测任务，需最大限度从语料中提取档案信息特征供算法模型使用，方能提高预测准确率，这对综合档案馆是繁重且具挑战性的工作。

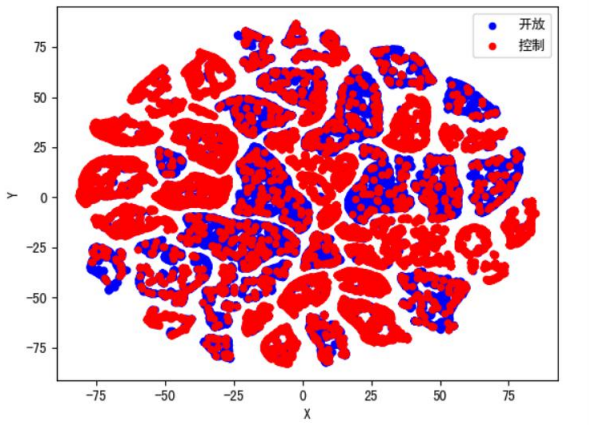

图 3 t-SNE 分布式随机邻域嵌入分析示意图

c.深度学习法。基于深度学习的方法可绕开特征工程，能较好实现线性或非线性数据的拟合，发现并刻画档案数据内部复杂的结构特征。为此，课题组采用深度学习框架 Tensorflow 作进一步研究，用 python 构建了一个含四层隐藏层的深度学习感知器二分类算法模型，将之前实验保留的多维度样本原始特征传给深度网络，由该模型自动计算权重系数用于二分类预测，但是，模型并不输出每一隐藏层形成的权重系数。运用该模型所做实验的预测准确率有进一步提升，3 次多维加权组合鉴定平均准确率为 67.85%，总准确率达到 73.16%。此外，经过对 sigmoid 函数、tanh 函数的比较实验，选用 sigmoid 函数计算深度学习感知器二

分类算法模型预测结果的置信度，作为向开放鉴定流程推荐预测结果强度的参数。

三、成果的创新点：

1.创新点

课题综合运用数据挖掘等人工智能技术、关键词匹配技术、档案知识服务和专家经验，创造性提出了从16个维度辅助档案开放审核的智能化方法，构建了由递进式辅助开放鉴定双模块、相关算法模型、深度学习神经网络模型以及档案知识库组成的辅助开放鉴定模型，形成可用的人工智能相关开源技术“工具箱”，提出了创建、维护与管理辅助开放鉴定模型的新方法。

2.保密要点

本课题研究成果中，除已经和今后公开发表的成果外，其它属于本课题的研究成果均属保密范围，包括课题研究报告、实验记录、算法模型、深度学习感知器二分类模型、全套应用系统设计文档与代码、档案知识库等。