面向数字记忆开发利用的档案检索模型构建研究*

房小可
（北京联合大学应用文理学院，北京 100191）

摘要：本文从构建目标、数字记忆开发利用途径和档案组织粒度三方面探讨数字记忆和档案检索之间的逻辑关系，并在此基础上构建面向数字记忆开发利用的档案检索模型。模型主要分为档案信息中数字记忆要素提取、要素语义关系提取和索引库建立及匹配。其中数字记忆要素提取分为基于命名实体识别方法的实体要素提取，以及基于LDA模型的主题提取；要素语义关系提取分为基于神经网络的实体关系提取和基于空间向量相似性的主题关系提取；索引库建立及匹配模块旨在通过检索数字记忆要素字段获取档案承载记忆的基因链，实现记忆的完整再现，促进档案价值的开发利用。

关键词：数字记忆；信息检索模型；数字记忆要素；语义关联
中图分类号：G270.7　DOI：10.3772/j.issn.1673-2286.2021.11.003
引文格式：房小可. 面向数字记忆开发利用的档案检索模型构建研究[J]. 数字图书馆论坛，2021 (11)：21-27.

自1994年美国记忆启动以来，数字记忆项目在全球范围生长蔓延，建设主体和主题类型丰富多样，很快形成繁茂景象[1]。不仅成立了国家记忆项目如美国记忆、中国记忆等，地区层面的记忆项目也层出不穷，如北京记忆、香港记忆等。更有泛在用户参与的面向数字记忆建构的档案著录工具ICA-AtoM（Access to memory）[2]。此外，在《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》[3]中，也强调要加快数字化发展，建设数字中国。档案学领域也开展了档案与社会记忆、城市记忆、国家记忆之间的探索，提出并论证了“档案记忆观”理论[4]。“档案记忆观”重要的内核之一是“档案具有记忆属性”[5]。在数字记忆已成为研究重点的当下，对档案承载的记忆进行构建和挖掘，从而实现数字记忆的开发利用是值得研究的问题。对数字记忆开发利用比较普遍的方式是构建数字记忆网站，如美国9·11数字档案网站、佛罗里达州记忆网站等。信息检索是档案开发利用的关键方法，档案检索程度和质量直接关系着档案的开发利用程度和质量[6]，这些记忆网站均有在线分类浏览和档案主题检索的功能，用户通过关键词检索可获得匹配的档案资源。然而，记忆是对过去的感知与再现，每个故事均有一定的结构特征，包括叙述者、情节、场景、人物、危机与结局，那些能够记忆并再现出来的经历会成为故事所叙述的基本内容[7]。在数字化转型时代背景下，针对大量的数字记忆仅以数字化的档案照片或文件的形式将检索记忆结果反馈给用户是不够的，需要针对档案所承载记忆的结构特征，细粒度挖掘记忆要素，实现细粒度化的档案检索。本文正是以此为研究切入点，将数字记忆网站中的检索模块单独提出，面向数字记忆的开发利用来构建档案检索模型，为数字记忆细粒度开发利用提供借鉴。

1 国内外研究现状

由于直接面向数字记忆开发利用而构建档案检索模型的研究比较少，本文拟分别从档案视角下数字记忆

* 本研究得到国家社会科学基金青年项目“面向社会记忆构建的档案资源检索研究”（编号：18CTQ041）资助。

开发利用和档案检索两部分来进行梳理。

1.1 档案视角下数字记忆的开发利用研究现状

数字记忆的开发利用从开发视角上可分为三方面：一是数字记忆开发路径研究，即从横向开发主体到纵向信息资源采集、整理及利用的整体视角提出数字记忆开发方案。例如：霍艳芳等[8]用数字人文的理念与方法重新审视城市记忆资源建设模式，提出从资源采集到资源数据库搭建的全流程来优化传统资源整合模式和开发路径；Mina等[9]认为数字转型下文化传统与城市生活密不可分，对此作者梳理当地文化技术方面的举措，并以欧洲数字图书馆为例，介绍文化数字化方面取得的主要成就。二是数字记忆平台建设研究，即从资源采集到开发整体流程为主线，实现档案开发利用的虚拟平台。例如：冯惠玲等[10]在数字记忆理念下，以浙江台州古村落为对象，对已有资料进行数字化采集、加工，实现“记·忆高迁”网站平台的建设；Spagnoli[11]认为“临时展览虚拟档案”项目涉及开发在线档案，从而能够记录、保存和提供与临时展览和文化活动设计有关的数字材料，由此可将虚拟档案馆和博物馆作为保存和记录虚拟档案的主体，从而在文化遗产价值等方面发挥作用。三是面向数字记忆开发利用的档案资源建设研究，包括对档案资源库的建设、档案异构数据整合及档案知识图谱开发等。例如：牛力等[12]从异构记忆资源整合对象、整合基础、整合思路与整合技术四方面对异构记忆资源整合的研究现状进行系统梳理并剖析当前问题，提出解决对策；Hsieh等[13]针对我国台湾地区体育事业，将体育界重要人物的珍贵文物档案数字化，并构建数据库，进而通过线上线下开发提供体育文化多样性展示。

以档案为视角的数字记忆开发利用，其研究更多是基于某些开发手段形成数字记忆的展示形式（如网站、展览等），而未对档案中记忆的故事性和叙事性内容进行深度挖掘，且对于数字记忆的进一步检索利用研究较少。

1.2 档案检索研究现状

2000年以后，档案检索集中在以下三方面。一是档案网站检索研究，如赵屹等[14]以美国网络档案检索系统ARC为例，从档案源、著录项、检索途径、检索新功能、系统数据及检索性能介绍NARA提供的检索工具。二是信息描述与元数据研究，例如：Riley等[15]讨论了可共享元数据原理及应用于档案描述所涉及的问题、工具和策略；王兰成[16]从语义视角研究基于语义的档案信息整合及基于XML、EAD异构档案信息组织及其本体方法的应用。三是档案检索系统研究，例如：赵雪芹[17]通过分析现行检索服务存在的弊端及用户面临检索困境，提出将资源发现服务作为一种高效便捷的资源揭示和检索系统；Ricardo[18]在基于可扩展标记语言EAC-CPF（编码档案上下文）基础上，提出用于档案信息系统的协作框架，该框架支持辅助导航和主题映射，并提供语义丰富的访问层以确保不同归档保存记录的位置，改善了用户与网络的交互体验方式。

上述研究可知，对档案检索的研究大体上是将档案视为一种普通信息资源来处理，但档案承载的记忆具有故事性。未对档案承载的记忆特征进行分析而直接实现检索服务，难以为用户提供精准的档案服务，影响档案价值的挖掘与传承。

总的来说，数字记忆开发利用及档案检索具有开发针对性不够明确、开发深度不足、展现形式缺乏细粒度化的问题。由此，本文针对数字记忆本身的特征，基于语义分析等方法，通过检索模型实现数字记忆的细粒度叙事型展现，支持数字记忆的深度开发利用，发挥档案所承载记忆的历史及文化价值。

2 数字记忆与档案检索的逻辑关系

2.1 数字记忆开发利用与档案检索模型存在目标统一性

信息检索模型是对文档和查询进行表示以及对它们之间的相关性进行描述的模型，实际上是为满足用户需求对信息资源进行重组而设计的一套匹配模式。因此，档案检索模型的构建目标是为了满足档案用户需求，从而促进档案价值的开发利用。数字记忆的概念最早由中国人民大学冯惠玲教授提出，其本身代表着数字技术和社会记忆的火花碰撞，随着社会数字转型，逐渐从成为社会记忆的主要形态。在冯惠玲教授所主持的“北京记忆”项目实践中，将其初步定义为应用数字技术对各种记忆资源进行数字化组织与再现，使之达到可解读、可保存、可关联、可再组、可传播与共享，进

而支持数字时代集体记忆的构建与传承[19]。可见，数字记忆开发利用的目标之一是对信息资源组织与再现从而满足用户的需求，实现共享利用。这种对于档案资源的重组以满足用户的需求，二者的目标具有统一性。

2.2　档案检索是数字记忆得以有效利用的途径

数字记忆是否得到有效利用与是否满足用户需求紧密相关。笔者在前期研究梳理中发现，档案界主要是以档案馆为中心参与社会记忆构建工作，通过编研、展览、拍摄视频等方式进行社会记忆的传播[20]。这些基本是从价值论层面来考虑数字记忆产品的提供利用问题，缺乏从需求论层面即直接从用户需求的角度探讨提供利用；档案检索是根据用户提出显性需求（如输入查询词等方式）为用户提供记忆资源，属于需求论层面范畴。因此，在数字社会的当下，档案检索不失为数字记忆有效利用的途径之一。需要进一步说明的是，随着档案数据化的不断深入发展，需要将档案进行数据化处理，即以数据为起点进行数字记忆构建及开发利用，对此也有学者提出基于数字人文视角的社会记忆构建[21]。从档案粒度上看，可构建档案数据化范畴下的档案检索模型以支持数字记忆的有效开发利用。

2.3　数字记忆的呈现方式影响档案检索模型的信息组织粒度

冯惠玲[1]根据记忆资源的呈现方式，将数字记忆粗略划分为展陈型和叙事型。展陈型主要是将一定专题的记忆进行系统化展示，以原生资源诉说记忆，体现为语义连续性。以往的数字记忆开发利用更多是此种展示形式。叙事型则主要是在该专题研究基础上，用数字资源体系化、逻辑化、叙述式地呈现客体记忆，可以是语义分散式的档案表达。由前文可知，档案检索是根据用户提出显性需求的方式为用户提供记忆资源，检索得到的记忆呈现形式既可以是客观展示的粗粒度全文展示模式，也可以是逻辑化呈现的细粒度可视化展示模式。

综上，数字记忆和档案检索模型存在目标统一、途径相通、互为影响的内在关联，因此从档案检索模型的角度呈现数字记忆并实现其开发利用是可行的，也是值得研究的。

3　数字记忆开发利用视角下的档案检索模型构建

由前文可知，信息检索模型有两个重要要素，即信息表示和相关性匹配；数字记忆基本分为展陈型和叙事型两种呈现方式。如何基于档案信息表示和相关性匹配实现展陈型和叙事型两种展现形式是本部分需要解决的问题。面向展陈型的档案检索与以往的检索无差别，即基于著录项目实现检索结果的某种次序展现即可；需要说明的是，面向叙事型的档案检索，由文献[7]可知，记忆是对过去的感知与再现，每个故事均有一定的结构特征，包括叙述者、情节、场景、人物、危机与结局，那些能够记忆并再现出来的经历会成为故事所叙述的基本内容。因此，呈现叙事型检索结果的前提不只是依据著录项目，更重要的是需要对档案承载的记忆进行记忆实体的挖掘和语义组织，即记忆要素的识别以及要素之间的关联构建。以此为依据，本文构建的面向数字记忆开发利用的档案检索模型如图1所示。其中面向展陈型的档案检索模块与当前档案检索模式基本一致，即将档案数字化并建立索引库，实现基于案卷名、文件名、文件形成时间等著录项目的检索，获取档案数字化副本。面向叙事型的档案检索模块构建的前提是将档案数据化，进而根据叙事特征和需求提取数字记忆的叙事要素及其语义关联，通过建立要素索引获取叙事网络，技术上实现语义检索，服务上还原事件的来龙去脉。依据模型拟解决的关键问题，本部分着重阐述面向叙事型的档案检索模块。

3.1　档案信息中数字记忆要素提取

笔者以往的研究中，已对社会记忆要素进行了分析和揭示，从历史题材角度提取故事基本内容的骨架元素，即为社会记忆要素[22]，其元素应包括时间、地点、人物、事件、主题五类要素。数字记忆作为社会记忆数字转型的主要形态，同样应包含这五类要素。时间要素和地点要素指该事件发生过程中出现的重要时间和重要地点；人物要素包括事件中出现的真实人物、组织团体或机构等；事件要素即一次活动或多次活动的集合，体现在案卷题名或文件题名中；主题要素是整个事件的重要故事节点。

数字记忆的各要素，时间、地点、人物、事件，从信息检索学科角度看均属于命名实体；而主题要素作为

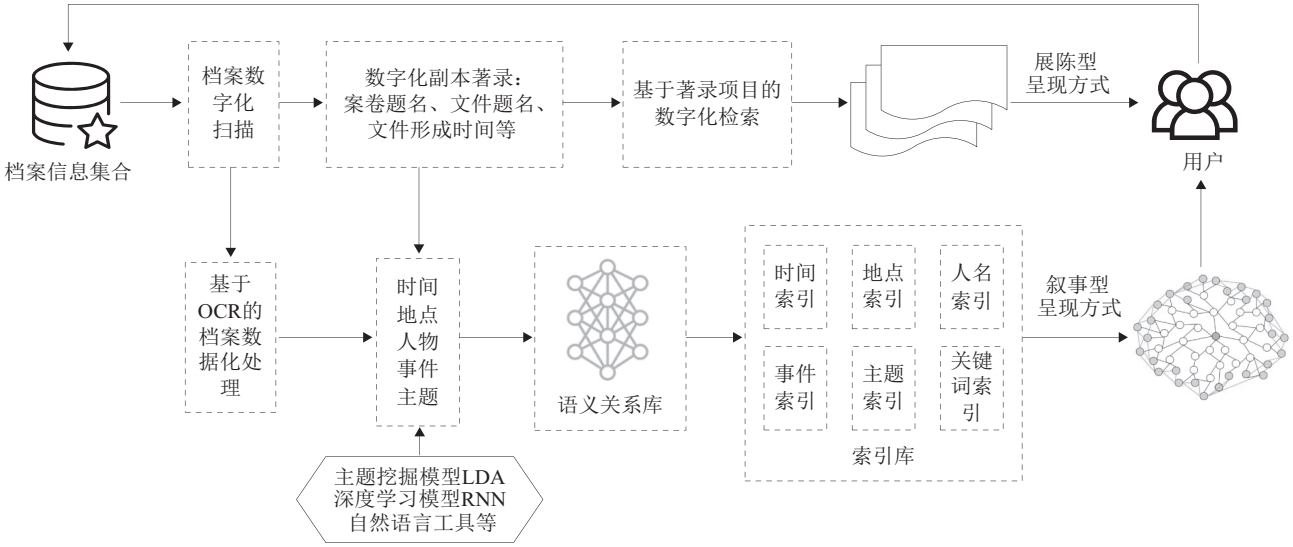

图1 数字记忆视角下档案检索模型

事件的主要内容可通过主题模型或聚类等方式提取。因此，数字记忆要素提取工作可转换为命名实体识别及主题挖掘两项任务。

3.1.1 基于规则方法的时间和地点命名实体识别

命名实体识别的方法主要分为两种，一是基于规则和字典的方法，二是基于统计的方法。基于规则的方法和基于字典的方法都是要构建大量的规则集或字典，然后按照需求将需要识别的汉字串放入制定的规则集中或与所构建的字典进行匹配，经过多次修正直到匹配成功。具有代表性的是Colllins等[23]提出先定义种子规则集Decision List，再根据语料对该集合进行无监督的训练迭代得到更多的规则，最终将规则集用于命名实体的分类。基于规则的实体识别比较适用于形势比较固定、规则比较容易提取的命名实体，如时间、地名。由此，档案数字记忆的时间和地点要素可基于规则的命名实体识别方法。

3.1.2 基于统计的人物和事件命名实体识别

基于统计的命名实体识别，目前比较有效的是序列化标注方法，即对于文本中每个词，可以有若干个候选的类别标签，这些标签对应其在各类命名实体中所处的位置，对其进行训练进而实现分类。如GU等[24]运用条件随机场和知识库，通过分析中文名字的特征，将中文人名的训练集进行序列标识，并对测试集进行命名实体识别。

档案信息中数字记忆的人物要素和事件要素与时间和地点实体不同，表达形式一般为自然语言，但是也有规律可循。对于人名，其上文一般是“称呼”“职衔”，下文一般是“先生”“同志”“说”之类的称谓词或动词，根据具体情况对数据集进行序列标识。以一份民国档案文件“里昂中法大学写给校董蔡元培先生的信件”作为分析，对于“蔡元培”这一人名要素名称，人名前是“职衔”的称呼，即“校董”，人名后是称谓词“先生”。如果对整个句子从人名内部组成、上下文、无关词进行标注，对整个句子进行标注的结果是：“里昂中法大学/RN 写/RN 给/RN 校董/RQ 蔡/RX 元/RM 培/RM 先生/RH 的/RN 信件/RN”，其中RN表示与人名无关的语词，RQ是人名上文的词，RX是人名的姓氏，RM是人名的名字，RH是人名后文的词。然后通过基于Viterbi算法的自动标注和识别得到人名实体。事件要素同理，这里不再赘述。因此档案数字记忆中人物要素和事件要素可考虑基于统计的命名实体识别方法。

3.1.3 主题要素的提取

Blei等[25]于2003年提出了隐含狄利克雷分配（Latent Dirichlet Allocation，LDA）模型，该模型对参数自身提出了先验假设，属于完全概率生成模型，因此是一个三层贝叶斯模型。与PLSA相同，LDA假设文档表示为

主题的概率分布，而主题表示为词语的概率分布，是目前应用广泛的模型之一。学者基于研究对象的不同，对LDA主题模型进行了拓展和改良，最具有代表性的改良LDA模型包括：基于ATM（AuthorTopic Model）的主题建模、Twitter-LDA主题建模和基于Labeled LDA的主题建模等。本文选用LDA模型实现主题挖掘。

基于LDA的档案信息主题挖掘的主要思想是，认为每份文件是若干主题的混合分布，而每个主题又是由若干词汇（包含命名实体）组成的概率分布。因此可以将每份文件表示为这些隐含主题的概率分布（file-topic），而每个隐含主题可表示为词汇的概率分布（topic-word）。主题要素丰富了记忆的叙事性，是数字记忆不可缺少的再现情境。

3.2 数字记忆要素语义关联抽取

数字记忆要素语义关联抽取实际上分为两种类型的关系抽取：一是实体关系抽取，如人物-地点、人物-事件，或者要素内部实体如人名-机构名、事件1-事件2等；二是主题之间的语义关系抽取。

对于实体关系抽取，已有的方法主要是从语料信息中提取词性、句法结构、语义依存关系等表面特征和结构化特征，并用模式匹配、特征向量和基于核函数的方法对实体对之间的关系进行分类[26]。这些实体关系抽取方法前期对自然语言处理工具具有较强的依赖性，因此受到自然语言工具处理结果的影响。深度学习的概念最早是在2006年由Hinton等[27]正式提出。基于深度学习的实体抽取方法能够自动提取特征，减少对人工的依赖，且具有良好的泛化能力，可用于抽取大规模文本数据。其中，CNN和RNN是实体关系抽取中应用比较广泛的网络模型，考虑两种模型对文本处理的效果，本文选择RNN作为实体关系抽取模型，并引入注意力机制为每个实体计算一个关系权重，以此提取数字记忆基因链，为后续数字记忆检索叙事化呈现提供数据支持。

对于主题之间的语义关系抽取，可根据向量之间的相似度抽取主题之间的语义强弱关系。由前文可知，每个主题由若干个有实际意义的词汇组成，若干词汇概率形成概率分布，因此每个主题可用一系列具备概率权值的词向量表示。运用主题向量之间的余弦相似性计算可得到每个主题之间的相似度值，值越大说明两个主题越相关，反之越不相关。

3.3 索引库建立及档案信息匹配

索引款目是有关信息资源所涉及的主题、事物及其他特征的信息单元，并指向其地址的一条记录[28]。因此，对于数字记忆中的索引库除了包含构建以原有著录项目中的关键词索引，还应构建数字记忆要素索引，即人物要素包含的人名索引、机构索引，以及其他要素中包含的时间索引、地名索引、事件索引和主题索引。索引地址指向与索引词具有语义关联的重要数字记忆要素，呈现实体语义关联，体现数字记忆基因链，还原事件来龙去脉和历史原貌。

模型根据用户输入查询词提取用户需求，形成布尔逻辑表达式，如果表达式提取后只对应一个语词，可以直接将提取的语词与索引进行匹配，一方面可得到以该语词为关键词的展陈型数字化档案；另一方面提取包含该语词的事件基因链。例如，在含有北京联合大学校址记忆的档案信息中，用户输入词为“北京联合大学应用文理学院”，则可基于关键词的全文检索获取该词所在的数字化文件或资料，得到展陈型的检索结果；进一步地，依据已有的事件的语义关联，获取从实体到实体的发展链条，如从“中国人民大学二分校”到“北京联合大学应用文理学院”的关系链条，以及在这一发展链条中所发生的历史故事的来龙去脉。

4 实证分析

本研究的实验部分以北京联合大学编著的《校址的故事》为数据源。该书在学校前党委书记韩宪洲的亲自指导下，由档案（校史）馆牵头编写，编写过程中小组成员不断挖掘馆藏档案，赴国家档案局、北京市档案局、北京市方志馆、平谷区档案馆等地查询确认每个信息点，历时2年多，记录了大学分校时期至今的校址变迁。以《校址的故事》为研究对象，通过本文提出的方法，即命名实体识别、实体关联挖掘还原事件的来龙去脉。为了清楚展示效果，采用微软开发的跨平台开放工具Visual Studio Code，选用jQuery作为优化HTML的辅助工具，其他前端可视化工具包括HTML、CSS、JS、Layer和G6。校址检索选择界面见图2。

图2中左边一栏中选择任何一个校址，即可显示介绍、时间线、主题、人名、机构名、关系图6个模块。介绍模块主要是对该校址的整体说明；时间线模块是对该校址的重要时间及其对应的事件进行梳理，可基于

规则的方式提取时间实体；人名、机构名模块分别基于统计的命名实体识别得到；主题模块基于LDA提取得到。

以“西城区西四丰盛胡同13号”为例，时间上从1978—2012年，共经历了9个校址，包括中国人民大学第二分校校址、北京联合大学文法学院院址等。基于LDA实现主题挖掘，经测试主题数为7效果较好：一是校舍的建立和设计，二是该地址所在校区硬件设施的建设，三是软件设施筹备，四是首次招生活动，五是专业设置，六是档案学专业的成立，最后是其他方面的支持，通过主题挖掘能较清晰的反映出有关该校址阐述的主要环节和内容；最终，通过实体之间的关系深化用户对该校址的理解，从显性的实体展示过渡到隐性的实体之间的关系。

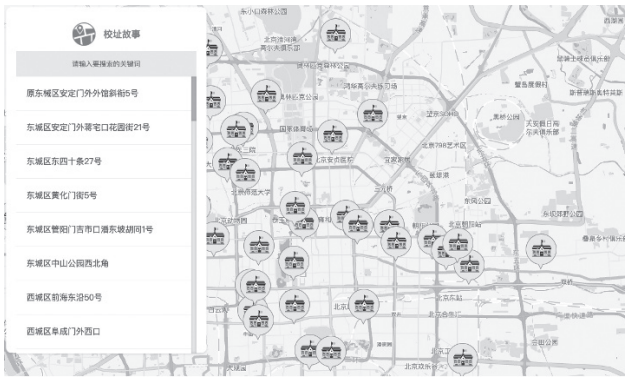

图2 校址检索选择界面

5 总结

当前的档案检索模型更多将档案作为普通信息进行处理，然而档案承载的记忆具有一定的结构特征，包括叙述者、情节、场景、人物、危机与结局等，因此需要针对记忆特征实现检索及记忆结果呈现。数字记忆的呈现方式主要有展陈型和叙事型两种方式，对于面向叙事型的数字记忆开发利用当前研究尚且不足。由此，本文剖析数字记忆与档案检索的逻辑关联，针对展陈型和叙事型两种呈现方式，构建面向数字记忆开发利用的档案检索模型，并细致阐述针对叙事型检索模型的构建过程和关键点。由于数据源的限制，本文的实证部分是对一次文献的二次开发和重组，未来的研究会增加数据量及不同档案数据类型，完善本文提出的方法。

参考文献

[1] 冯惠玲. 数字记忆：文化记忆的数字宫殿 [J]. 中国图书馆学报，2020，46（3）：4-16.

[2] ICA. ICA-AtoM Project [EB/OL]. [2021-11-01]. https://www.accesstomemory. org/en/docs/2.7/.

[3] 新华社. 中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要 [EB/OL]. [2021-11-01]. http://www.gov.cn/xinwen/2021-03/13/content_5592681.htm.

[4] 龙家庆，聂云霞. 数字记忆建构视域下档案文化创意服务模式探析 [J]. 档案学通讯，2020（5）：68-76.

[5] 加小双，徐拥军. 中国“城市记忆”理论与实践述评 [J]. 档案学研究，2014（1）：22-32.

[6] 于力春，李健. 语义检索——档案智能检索的新方向 [C] //2019年全国青年档案学术论坛论文集. 重庆，2019：35-41.

[7] 丁华东. 昔日重现：论档案建构社会记忆的机制 [J]. 档案学研究，2014（5）：29-34.

[8] 霍艳芳，何思源. 数字人文视阈下城市记忆资源整合与开发路径研究 [J]. 档案学研究，2018（1）：29-34.

[9] MINA F I，PANA M C. From culture to smart culture. How digital transformations enhance citizens’ well-being through better cultural accessibility and inclusion [J]. IEEE Access，2020，8：37988-38000.

[10] 冯惠玲，梁继红，马林青. 台州古村落数字记忆平台建设研究——以高迁古村为例 [J]. 中国档案，2019（5）：74-75.

[11] SPAGNOLI A. Virtual archive of temporary exhibitions：New scenarios for the documentation，storage and fruition of an“ephemeral memory” [C] //Proceedings of the 2012 18th International Conference on Virtual Systems and Multimedia，Italy：Milan，2012：529-532.

[12] 牛力，赵迪，韩小汀. “数字记忆”背景下异构数据资源整合研究探析 [J]. 档案学研究，2018（6）：52-58.

[13] HSIEH Y G，CHEN T H. Digital archive use in physical education and sports culture [C] //ARCHIVING 2019：Digitization，Preservation，and Access - Final Program and Proceedings，United States，2019：120-124.

[14] 赵屹，陈晓辉. 美国网络档案信息检索系统ARC [J]. 北京档案，2003（7）：44-45.

[15] RILEY J，SHEPHERD K. A brave new world：Archivists and shareable descriptive metadata [J]. The American Archivist，2009（72）：91-112.

[16] 王兰成. 论实现异构档案信息整合的信息组织与检索技术 [J].

档案学研究，2011（2）：55-59.

[17] 赵雪芹. 档案数字资源发现服务研究 [J]. 档案学通讯，2013（1）：43-47.

[18] RICARDO E B. Context-based aggregation of archival data：the role of authority records in the semantic landscape [J]. Archival Science，2015，15（3）：217-238.

[19] 加小双. 档案学与数字人文：档案观的脱节与共生 [J]. 图书馆论坛，2019，39（5）：10-16.

[20] 房小可. 档案学科视角下社会记忆构建框架研究 [J]. 档案学研究，2021（3）：18-23.

[21] 杨文. 数字人文视阈下的社会记忆构建研究 [J]. 情报资料工作，2019，40（5）：38-45.

[22] 房小可，王巧玲. 档案著录、知识关联与社会记忆重构 [J]. 档案学通讯，2021（3）：27-33.

[23] COLLINS M，SINGER Y. Unsupervised models for named entity classification [C] //Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, Florham park，1999：100-110.

[24] GU C, TIAN X P, YU J D. Automatic recognition of Chinese personal name using conditional random field and knowledge base [J]. Mathematical Problems in Engineering，2015：1-6.

[25] BLEI D，NG A，Jordan M. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research，2003（3）：993-1022.

[26] 马语丹，赵义，金婧，等. 结合实体共现信息与句子语义特征的关系抽取方法 [J]. 中国科学，2018，48（11）：1533-1545.

[27] HINTON G E，SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks [J]. Science，2006，313（5786）：504-507.

[28] 马费成. 信息资源开发与管理 [M]. 北京：电子工业出版社，2014：128.

作者简介

房小可，女，1987年生，博士，副教授，研究方向：知识组织与知识服务、档案信息资源开发利用，E-mail：xiaoke@buu.edu.cn。

Research on the Construction of Archives Retrieval Model for the Digital Memory Development and Utilization

FANG XiaoKe
(School of Applied Arts and Science, Beijing Union University, Beijing 100191, P.R.China)

Abstract: This paper discusses the logical relationship between digital memory and archives retrieval model from the aspects of the goal of construction, the way of digital memory utilization and the granularity of archives organization. Then, this pater constructs archives retrieval model oriented to digital memory utilization. The model is mainly divided into three parts: the extraction of digital memory elements, the extraction of semantic relationship of elements, and the establishment of index database and resource matching. Digital memory feature extraction includes entity feature extraction based on named entity recognition method and topic extraction based on LDA model; Feature semantic relation extraction includes entity relation extraction based on neural network and topic relation extraction based on spatial vector similarity; The index database building and matching module aims to retrieve the digital memory element fields to obtain the gene chain of memory carried by archives, which realizes the complete reproduction of memory, and promotes the development and utilization of value of archives.

Keywords: Digital Memory; Information Retrieval Model; Elements of Digital Memory; Semantic Relevance

（收稿日期：2021-10-23）