

基于情景相似度的口述历史 资源事件自动识别研究^{*}

刘宁静 孙 翌 周 锋 刘 音

【摘 要】口述历史资源是个人记忆与社会记忆结合的一种资源形式，其形成过程也是社会记忆再生产的过程，具有典型的叙事特征。对口述历史资源中的名人历史事件进行语义建模，并在元数据层面开展名人历史事件识别具有重要的理论意义与实践意义。文章在构建语义模型的基础上，优化一种计算口述历史资源事件情景相似度的方法，以李政道和CUSPEA事件为例，构建实例事件的情景单元，提取情景要素后对李政道特藏资源中的口述历史资源进行实验。实验结果表明，在设置阈值为0.6时，判断准确率为76.19%，能较大提高此类资源事件标识的效率。

【关键词】口述历史资源 情景相似度 历史事件 自动识别

Abstract: Oral history resources are a form of resource that combines personal memory with social memory. The formation of these resources is also a process of social memory reproduction, characterized by distinctive narrative features. Semantic modeling of celebrity historical events within these archives, and the identification of such events at the metadata level, are of significant theoretical and practical importance. This paper optimizes a method for calculating situational similarity of event scenarios in oral history resources based on the construction of a semantic model. Taking the cases of Tsung Dao Lee and the CUSPEA event as examples, it constructs situational units for instance events and extracts situational elements for experiments with the oral history resources in Tsung Dao Lee's special collection. The experimental results show that with a threshold set at 0.6, the accuracy rate is 76.19%, significantly enhancing the efficiency of event identification in such resources.

Key words: oral history resources situational similarity historical events automatic recognition

DOI:10.15941/j.cnki.issn1001-0424.2024.09.004

0 引言

特藏是一个图书馆独具特色的重要馆藏，被视为图书馆的特殊标签。近年来，图书馆作为重要的社会记忆机构，越来越重视对承载社会记忆的口述历史资源的收集，此类资源也已成为图书馆特藏资源的一个重要组成部分。特藏资源智慧化建设是智慧图书馆建设的关键内容。当前，图书馆特藏资源的服务呈现出由“以藏为主”转向“藏用结合”的大趋势。图书馆除了需要保持对特藏收藏的关注之外，还应投入更多精力开展特藏的揭示、保护、服务、研究和展示，立体化、全方位地提供特藏服务^[1]。而要想革新现有服务为智慧知识服务，在特藏资源数字化的基础上，对其进行数据化加工、语义化处理、知识化组织和可视化呈现，是重中之重。

口述历史资源是以人物和事件为核心建设而成的专题性档案资源，其形成过程也是社会记忆再生产的过程，具有典型的叙事特征^[2]。以“发现”“重构”和“故事化”三原语为主线的档案研究已成为新的研究范式，主要通过将档案数据进行多维组织、梳理成册、多元可视、空间映射等方式，以主题故事化的形式抽取档案数据，对档案记忆进行叙事化表达，从而形成档案数据的故事化^[3]。在口述历史资源中，围绕名人开展的口述历史资源是

^{*} 本文系国家社会科学基金项目一般课题“数字人文视域下口述历史档案资源知识图谱构建及应用研究”（项目编号：22BTQ083）的研究成果之一。

常见的专题类型,本研究围绕此类典型口述历史资源展开。这类数字记忆具有重要的史料价值和精神传播价值,其中,人物是此类资源收集的核心标准,是揭示的核心要求,是管理和服务的核心元素。名人和与之关联的重要历史事件相辅相成,不可分割。一方面,关键人物往往对与之关联的重要事件发挥着决定性的作用;另一方面,关键人物因为某一或某些重要事件的历史作用、重要影响较高知名度,而成为名人。可以说,事件是围绕名人开展的口述历史资源所不可或缺又至关重要的语义单元,对此类口述历史资源进行相关历史事件的主题识别与标引是对资源进行语义化处理和知识化组织的基础工作。

当前,图书馆对口述历史资源进行历史事件识别与标引的方式主要有两种。一种是人工识别著录,如上海交通大学李政道图书馆所藏口述历史资源,将资料分为28个主题,并由专业馆员对资料进行人工识别与分类著录。这种方式存在明显的局限性:第一,人工识别著录较耗费时间和人力,对制定主题的人员具有较高的要求;第二,人工识别著录质量难以保证,由于参与标注的工作人员较多,对名人及相关事件的认知程度不一,识别和标注的质量参差不齐,存在漏标和错标的现象,对后续资源的使用有较大影响;第三,人工识别著录的可扩展性不强,例如随着研究的深入和用户需求的变化,人工预定义的主题常常出现不敷使用和急需增加的窘况。第二种是自动识别标引,通常使用BERT等预训练模型进行标引。这类模型比较依赖大量自然语言的训练,对于只有元数据、未进行过OCR识别的“小数据”文本则适用性不足。因此,对名人历史事件进行元数据级的语义建模,开展基于情景相似度的口述历史资源中事件自动识别和标引研究,具有重要的理论意义和实践意义。

1 国内外研究综述

对于情景,国内外学者进行了很多深入的研究。在突发事件预测领域,Kahn H和Wiener A^[4]于1967年最早提出情景的概念,认为情景是事物未来发展结果以及通向该种结果的途径的集合。情景就是以过去的经验描述关于未来的故事。情景要素是情景的最小构成单元,以情景要素为依据可对情景之间的关系进行分析,同时也能反映事件发生发展的状态与趋势^[5]。在自然语言理解领域,对事件情景有着不同的理解。晋耀红等^[6]认为情景是语境框架的一部分,语境框架把信息抽象成领域、情景、背景这3个侧面。其中领域是描述事件核心所归属的范围;情景是事件的动态描述,主要包括事件的内容、事件的参与者和事件参与者间的语义关系;背景包括语言中的褒贬倾向、陈述中心时空等上下文背景信息。王石颖^[7]认为,情景是用框架理论对事件进行知识表示的一种方法,建立事件框架可以对事件的不同方面以及事件之间的不同关系进行表述。在用户检索及个性推荐领域,学者们从知识本体、用户情感感知、语义分析等不同方面^[8],讨论了如何对用户所处的情景特征进行高效分析并归类,以达到更精准有效地进行关联检索或信息推荐。关于事件情景的构建,学界采取的方法和理论主要有六元素法(Event、SEM、ABC、LODE、CIDOC CRM、Event-Model-F)、三元组法、框架理论、知识元理论等。情景相似度是情景分析中的重要基础,也是情景分析能够具有更高普适性的重要因素。如:杨君等^[9]考虑重要的情景维度,构建基于情景相似度的多维信息推荐系统模型;张明红等^[10]横向上依据事件发生的顺序构建情景链,纵向上提出结构化突发事件模型,提出情景链相似度算法解决相似突发事件问题;徐绪堪等^[11]利用情景相似度,构建突发事件多粒度响应模型。在事件的识别标引方面:陈健瑶等^[12]从事件文本句法特征视角出发,提出一种在突发自然灾害领域,可较精准进行网络舆情具体事件识别的方法;蒋小霞等^[13]提出了一种在事件表示和对比学习的基础上对文本数据集进行深度聚类的方法,有效提高了聚类准确率。在事件识别方面:付剑锋等^[14]、王红斌等^[15]聚焦触发词,通过对触发词的依存分析研究词间关系,提取事件特征;张亚军等^[16]把事件识别作为分类问题,提出一种基于深度学习的事件识别模型,对触发词的识别问题转换为特征分类问题进行识别;陈健瑶等^[12]面向小规模数据集,从事件文本句法特征出发,提出一种对突然自然灾害网络舆情事件识别方法;王彦莹等^[17]针对历史古籍中的通用事件,构建了事件识别模型,相比基础的生成模型,识别效果大为提高。

总结来说,情景分析方法和情景相似度计算在突发事件的应急管理和用户信息推荐方面应用较为广泛。当前研究者采用情景应对模式(包括情景分层、情景推理、情景重建等)进行突发事件应急管理已基本形成共识,而基于情景相似度构建多维信息推荐模型进行用户信息推荐也日渐成为研究热点。但是,将其应用于口述历史资源的名人历史事件识别标引方面的研究还较为少见。故此,本文力求基于情景分析方法和情景相似度计算,探索一种对口述历史资源事件进行识别和标引的方法。

2 情景相似度计算

本节在构建名人历史事件语义模型的基础上,对具体的口述历史资源和名人历史事件进行实例化后构建情景单元、提取要素并进行语义化扩展,最后通过情景相似度分析对历史事件进行识别和标引。

2.1 语义模型构建

人、地、时、事、物是哲学层面描述客观世界的5个基本概念,是理解现实世界的基本语义单元,同时也是挖掘口述历史资源语义知识的基本实体^[19]。刘宁静等^[18]在综合研究前人语义模型的基础上,运用自上而下的方法,结合名人口述历史资源的特殊性和应用需求,构建而成的名人历史事件语义模型如图1所示。该模型主要包括时间、人物、资源、空间、描述(关系)等要素。该框架也是后续提炼情景单元的基础。

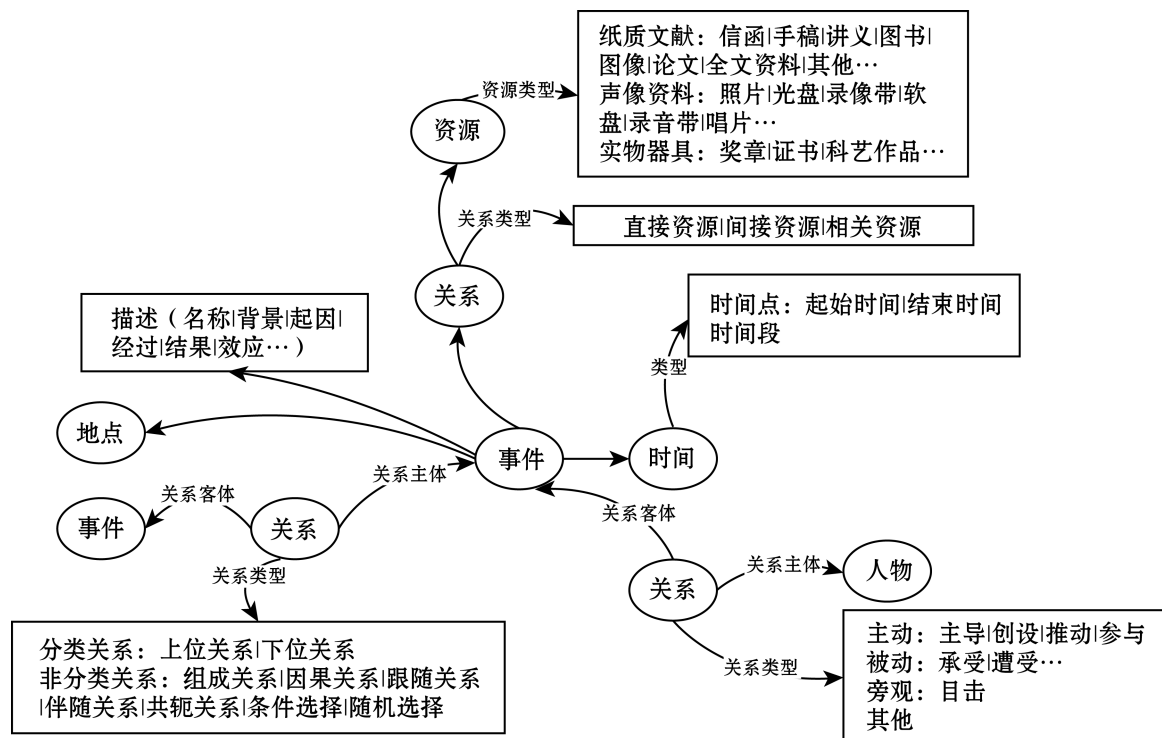


图1 名人历史事件语义模型^[18]

作为科学家名人的李政道先生一生致力于祖国的科技教育事业。1979年,李政道先生以战略科学家的眼光创设中国-美国联合招考物理研究生项目(China-U. S. Physics Examination and Application Program, CUSPEA),并身体力行地推行了十年,培养了915位CUSPEA学者,很多人成为学界和业界的领军人物。对于李政道特色资源来说,CUSPEA是一个资源主题,也是一个历史事件,而将与CUSPEA事件相关的资料标引出来,是历史研究和精神传播需要的基础。本文以李政道和CUSPEA为例,以馆藏李政道特色资源中CUSPEA相关资源为基础数据,对名人历史事件语义模型进行实例化,构建情景单元,提取要素并进行语义化扩展,计算其情景相似度。最后根据人工判断结果计算准确率。

2.2 情景单元构建和要素提取

关于事件情景的构建,学界采取的方法和理论主要有六元素法(Event、SEM、ABC、LODE、CIDOC CRM、Event-Model-F)、三元组法、框架理论、知识元理论等。

曲毅等^[19]借鉴共性知识元模型,从空间、时间及对象3个方面界定情景构建模型。文斌等^[20]基于对突发公共事件的层次分析,将应急管理涉及的事物系统分为客观事物、突发事件、应急管理活动3个子系统,并结合共性知识元模型,详细构建了突发事件情景中客观事物、突发事件、应急活动三类元素的知识元表示模型。HNC理论认为事件的语境框架由领域、情景、背景组成,其中情景可表示为 $SIT = SCD(A, B, C)$ 。近年来,使用知识元理论构建事件情景逐渐成为主流方法。

本文在人物历史事件框架下,将CUSPEA事件用知识元的方法,提取情景单元S,表示为:

$$S = (A, P, T, E)$$

①

其中A为CUSPEA关系人集($A_1, A_2, \dots, A_n$),包括CUSPEA学者、CUSPEA相关机构和参与过CUSPEA项目的工作人员等。P为CUSPEA相关地点集($P_1, P_2, \dots, P_n$),例如历年举行CUSPEA考试的考场所在地,CUSPEA办公室的所在地等。T为CUSPEA事件中的重要时间节点或时间段集($T_1, T_2, \dots, T_n$),例如CUSPEA

存续期 1979—1988 等。E 为 CUSPEA 子事件集 (E1, E2, …, En), 例如准备、考试、阅卷、申请、录取、入学等。

2.3 情景要素的语义化扩展

基于情景框架, 本文以 CUSPEA 事件代表性文本进行语义特征词提取。

CUSPEA 事件代表性文本主要指经过专家筛选和审定与 CUSPEA 相关的资料, 主要包括介绍 CUSPEA 项目、CUSPEA 学者的相关口述历史资料, 例如《CUSPEA 十年 (第二版)》《我和 CUSPEA》等 CUSPEA 学者口述历史采访资料, 以及李政道特藏资源中以 CUSPEA 为关键词查找所得藏品的元数据等。

本文在代表性文本的基础上使用关键词分析辅助专家知识判断, 进行 (A, P, T, E) 数据集的整理, 详见图 2。

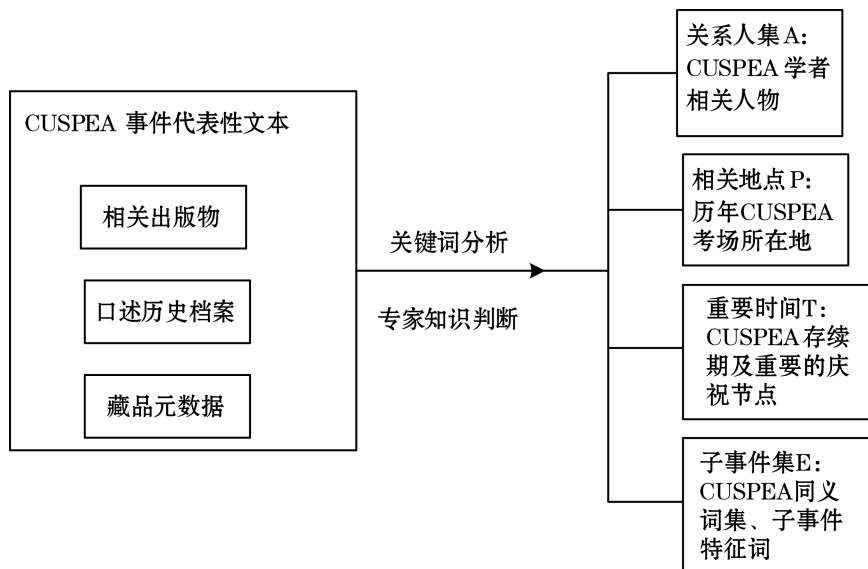


图2 情景要素的语义化扩展整理

关系人集 A 方面, 整理 CUSPEA 学者名单 915 人, 参与过 CUSPEA 项目的相关人物 375 人 (包括 CUSPEA 办公室工作人员、CUSPEA 委员会成员、参加物理阅卷的教师、参加英语阅卷的教师、来华面试的美国教授夫妇、支持 CUSPEA 项目的政府官员等), 相关机构 155 个 (包括 CUSPEA 委员会专家机构、参加 CUSPEA 项目的中美大学和研究机构)。

相关地点集 P 方面, 整理历年举行 CUSPEA 考试的考场所在地 17 个。

重要时间集 T 主要为 CUSPEA 存续期和重要的庆祝节点, 如 CUSPEA 二十周年、四十周年等。

CUSPEA 子事件集 E 方面, 主要整理了 17 个 CUSPEA 的同义词集和子事件的语义特征词。

2.4 情景相似度分析

在突发事件应急处理研究领域, 比较经典的情景相似度计算方法是通过对源案例与目标案例 (新案例) 所包含的特征属性, 计算两者交集和并集的权重之和的方法来计算情景的结构^[21]相似度, 如式②所示。

$$\varphi(A, B) = \frac{W_{A \cap B}}{W_M} = \frac{\sum_{i=1}^m w_i}{\sum_{k=1}^l w_k} \quad (2)$$

式②中: A 为目标案例情景的非空属性集; B 为源案例情景的非空属性集; w_i 表示案例 A 和案例 B 交集的第 i 个属性的权重; w_k 表示案例 A 和案例 B 并集的第 k 个属性的权重^[21]。

本文所研究的名人历史事件领域本质上都属于同一历史事件, 并没有严格的源案例与目标案例之分, 如果使用突发事件案例的相似度计算来类比, 可以将已知确定为某一历史事件的材料作为源案例, 将未知历史事件归属的待标注材料作为目标案例 A。在这种情况下, W_M (即源案例与目标案例的非空属性并集) 不变, 仅需比较 $W_{A \cap B}$, 式②可简化为式③。

$$\varphi(A, B) = W_{A \cap B} = \sum_{i=1}^m w_i \quad (3)$$

3 应用实践与分析

3.1 实验设计

(1) 实验数据集。实验所用数据集是李政道图书馆所收藏的李政道特藏资源中口述历史资源相关的图像、图

书、信函类资源,以式(3)的方法为依据进行相似度计算。选择与CUSPEA项目时间重合的档案箱B002为研究对象,取信函类资料作为待人工判断处理的实验语料。该箱资料共有信函1398件,此前已知确定与CUSPEA相关的信函数量为1,其余1397件信函资料与CUSPEA的相关性未知。

(2) 属性权重。对情景要素进行语义化扩展后的(A, P, T, E)数据集即为CUSPEA事件属性值,通过专家知识根据一定规则对其赋予权重,权重取值范围为[1~10]。

(3) 算法设计。使用情景相似度算法式(3),每次读取实验语料中的一条记录,依次检索记录中有内容价值的29个字段中是否包含属性值,如果包含则记录其权值,属性权值之和为该记录的相似度,使用最大最小标准化方法对该值进行归一化处理,形成归一化处理后的相似度。

3.2 实验结果

李政道特色资源B002箱共有信函1397件,其属性权值之和范围为[0, 67],对其进行归一化处理后设置阈值为0.6,则有42件信函相似度大于0.6,其余1355件信函相似度小于0.6。对相似度排名前100的信函逐一进行人工判断,限于篇幅所限,部分应用实验结果如表1所示。

表1 部分应用实验结果示例

序号	分类号	权重总值	归一化相似度	A	P	T	E	人工判断	人工判断赋值
1	C002050104	67	1.00	47	20	0	0	是	1
2	C002050102	66	0.99	46	20	0	0	是	1
3	C002170105	56	0.84	26	20	0	10	有关	0.7
4	C002050047	53	0.79	33	20	0	0	是	1
5	C002050107	53	0.79	33	20	0	0	是	1
6	C002130159	53	0.79	18	20	0	15	是	1
7	C002190091	51	0.76	31	20	0	0	是	1
8	C002190111	50	0.75	30	20	0	0	是	1
9	C002130030	48	0.72	18	20	0	10	是	1
10	C002130082	48	0.72	13	20	0	15	是	1
11	C002030053	46	0.69	26	20	0	0	有关	0.7
12	C002030054	46	0.69	26	20	0	0	有关	0.7
13	C002030060	46	0.69	26	20	0	0	是	1
14	C002050036	46	0.69	26	20	0	0	是	1
15	C002050041	46	0.69	26	20	0	0	是	1
16	C002050049	46	0.69	26	20	0	0	是	1
17	C002050051	46	0.69	26	20	0	0	是	1
18	C002050053	46	0.69	26	20	0	0	是	1
19	C002050055	46	0.69	26	20	0	0	是	1
20	C002100076	46	0.69	26	20	0	0	否	0
……(限于篇幅有限,后续省略)									

对信函逐一阅读,进行人工判断有3种结果,其中“是”为该信函是CUSPEA事件直接相关信函,人工判断赋值为1;“有关”为该信函中未提及CUSPEA事件,但通信人员为参与CUSPEA的阅卷人员等相关人员,人工判断赋值为0.7;“否”为该信函与CUSPEA事件不相关,人工判断赋值为0。人工阅读排名前100的信函后可得,归一化相似度 ≥ 0.6 的信函中,判断准确率为76.19%(见图3)。

3.3 成效分析

口述历史资源内容上具有以人物、事件和主题为中心的典型特征,尤其关注个体的记忆、经历、体验和感受^[22]。当前,档案馆和图书馆等所藏的口述历史资源,大多参考元数据方案或工作规范进行内容标引,例如《口述历史电子档案元数据方案》(DA/T-2015)和《口述史料采集与管理规范》(DA/T 59-2017)等,口述历史资

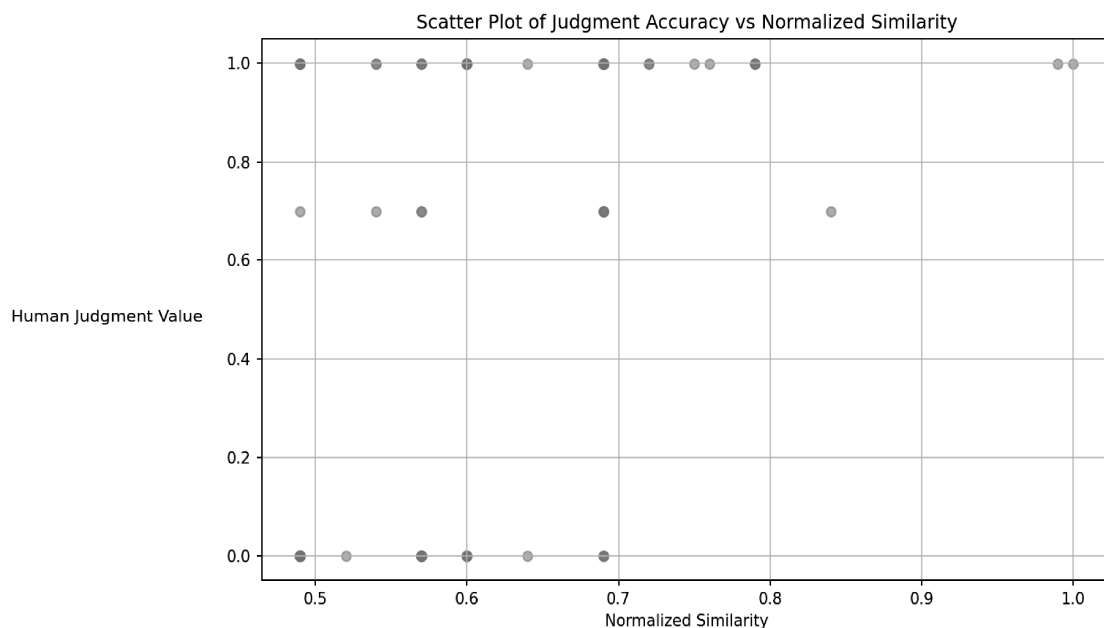


图3 情景相似度与人工判断散点图

源内的事件实体却无法被识别和利用。

本文所提出的以人、地、时、事为框架的情景相似度计算方法，对当前口述历史资源现状具有良好的适应性，促进诸多有益的转变。

(1) 文本资源向智慧数据转变。通过实验结果来看，76.19%的准确率可以在一定程度上实现事件的知识表示，即实现对口述历史资源的内容规范化表征^[23]，建设机器可理解的叙事性文本语料库，促使口述历史资源的叙事文本向智慧数据转变^[24]。

(2) 人工著录向智能标引转变。口述历史情节的描述需要深入理解事件的意义与组成，以李政道图书馆收录的CUSPEA主题事件为例，通过测试与推算，在B002箱1398封信函中，CUSPEA事件识别量由1提升为32，其人工判断与标引的工作时长至少需要3个星期，而智能标引则可进行协同过滤，快速定位且推荐关联事件，大大提高了事件实体的识别效率。

4 结语

本文在前人研究和构建情景化的名人历史事件语义模型的基础上，以科学家李政道和其创办的CUSPEA事件为例，以馆藏李政道特色资源中CUSPEA相关资源为基础数据，对名人历史事件语义模型进行实例化，构建情景单元，提取要素并进行语义化扩展，计算其情景相似度进行语义识别。经人工逐一判断，设置阈值为0.6，自动识别的准确率为76.19%。然而，由于时间、精力有限，本课题还有诸多不足之处，需在后续研究中继续深入。具体而言：(1) 特征属性权重采用专家知识赋值，主观性较大，噪音较为严重，后续拟使用AIGC等人工智能相关方法进行深入研究；(2) 实验范围较小，仅检测了全部馆藏资源中的一箱一类资源，还需要扩大实验范围进一步验证；(3) 仅对名人历史事件进行了主题识别，对(A, P, T, E)数据集的关键信息未做详细深入标引。

当前学界和业界对口述历史资料中名人历史事件识别标引的研究较少，多采用人工的方式进行识别著录，无法满足日益增长的知识服务需求。本文对名人特藏资源中以口述历史资源事件为主题的自动识别标引探索了一种可行的方法，提高了识别效率，具有一定的理论和现实意义。

参考文献

- [1] 王乐. 略论高校图书馆特色馆藏建设的价值与发展方向[J]. 大学图书馆学报, 2020 (3): 12-17.
- [2] 顾静, 丁华东. 口述历史档案社会记忆再生产的基本特点与思考[J]. 北京档案, 2023 (6): 19-23.
- [3] 牛力, 高晨翔, 张宇锋, 等. 发现、重构与故事化: 数字人文视角下档案研究的路径与方法[J]. 中国图书馆学报, 2021 (1): 88-107.
- [4] Kahn H, Wiener A. The Year 2000: A Framework for Speculation on the Next Thirty- Three Years [M]. New York: MacMillan, 1967: 6.
- [5] 杨峰, 张月琴, 姚乐野. 基于情景相似度的突发事件情报感知实现方法[J]. 情报学报, 2019 (5): 525-533.

- [6] 晋耀红. 基于语境框架的文本相似度计算[J]. 计算机工程与应用, 2004 (16): 36-39.
- [7] 王石颖. 基于情景框架和依存树的突发事件相似度计算[D]. 大连: 大连理工大学, 2013.
- [8] 李枫林, 陈德鑫, 梁少星. 基于语义关联和情景感知的个性化推荐方法研究[J]. 情报杂志, 2015 (10): 189-195.
- [9] 杨君, 吴菊华, 艾丹祥. 一种基于情景相似度的多维信息推荐新方法研究[J]. 情报学报, 2013 (3): 262-269.
- [10] 张明红, 余廉, 耿波. 基于情景的结构化突发事件相似度研究[J]. 中国管理科学, 2017 (1): 151-159.
- [11] 徐绪堪, 李一铭. 基于情景相似度的突发事件多粒度响应模式研究[J]. 情报科学, 2021 (2): 18-23+43.
- [12] 陈健瑶, 夏立新, 舒怡娴. 基于句法特征的突发自然灾害网络舆情事件识别方法研究[J]. 现代情报, 2022 (6): 17-26+93.
- [13] 蒋小霞, 黄瑞章, 白瑞娜, 等. 基于事件表示和对比学习的深度事件聚类方法[J]. 计算机应用, 2024 (6): 1734-1742.
- [14] 付剑锋, 刘宗田, 付雪峰, 等. 基于依存分析的事件识别[J]. 计算机科学, 2009 (11): 217-219.
- [15] 王红斌, 郇洪奎. 基于 word2vec 和依存分析的事件识别研究[J]. 软件, 2017 (6): 62-65.
- [16] 张亚军, 刘宗田, 周文. 基于深度信念网络的事件识别[J]. 电子学报, 2017 (6): 1415-1423.
- [17] 王彦莹, 王昊, 朱惠, 等. 基于文本生成技术的历史古籍事件识别模型构建研究[J]. 图书情报工作, 2023 (3): 119-130.
- [18] 刘宁静, 孙翌, 刘音, 等. 基于口述历史资源的名人历史事件语义模型构建及实证研究[J/OL]. 现代情报, 1-12 [2024-04-28].
http://kns.cnki.net/kcms/detail/22.1182.g3.20240402.1528.002.html.
- [19] 曲毅, 仲秋雁, 马晓霏, 等. 基于情景的突发事件演变概率规则推理方法[J]. 系统工程学报, 2014 (4): 571-578.
- [20] 文斌, 张瑞新, 袁江南, 等. 突发事件应急管理知识元模型构建研究[J]. 数学的实践与认识, 2013 (17): 89-97.
- [21] 张炎亮, 徐元元, 李俊雅. 不完备信息下基于案例推理的城市火灾应急决策方法研究[J]. 中国安全生产科学技术, 2018 (3): 13-18.
- [22] 潘玉民, 吕芳. 口述档案资源建设探略[J]. 图书情报研究, 2010 (1): 51-54.
- [23] 宋宇远, 王晓光. 基于情节本体的叙事性文本语义结构化表示方法研究[J]. 中国图书馆学报, 2020 (2): 96-113.
- [24] 曾蕾, 王晓光, 范炜. 文档博领域的智慧数据及其在数字人文研究中的角色[J]. 中国图书馆学报, 2018 (1): 17-34.

刘宁静 上海交通大学图书馆李政道与特藏服务中心副主任, 馆员, 博士。研究方向: 知识组织和链接、特藏资源整理与挖掘。

孙翌 上海交通大学图书馆系统技术与空间服务中心主任, 副研究馆员, 硕士。研究方向: 数字人文技术、知识图谱。通讯作者。

周锋 上海交通大学图书馆馆员, 硕士。研究方向: 数字人文方法与技术。

刘音 上海交通大学图书馆馆员, 硕士。研究方向: 科学史。

(上接第 86 页)

- [25] PANDORA. Introduction [EB/OL]. [2023-10-21]. <http://pandora.nla.gov.au/>.
- [26] BnF. Archives De l'internet Site Institutionnel [EB/OL]. [2023-10-21]. <https://www.bnf.fr/fr/archives-de-linternet>.
- [27] Webarchive. lu. Luxembourg Web Archive [EB/OL]. [2023-10-21]. <https://www.webarchive.lu/>.
- [28] Archive-It. University of Texas at San Antonio Libraries Special Collections [EB/OL]. [2023-10-21]. <https://www.archive-it.org/organizations/318>.
- [29] The Swiss National Library. Web Archive Switzerland [EB/OL]. [2023-10-21]. <https://www.nb.admin.ch/snl/en/home/information-professionals/e-helvetica/web-archive-switzerland.html>.
- [30] IIPC. Legal deposit [EB/OL]. [2023-10-21]. <https://netpreserve.org/web-archiving/legal-deposit/>.
- [31] Narodna in Univerzitetna Knjižnica. Spletni Arhiv NUK [EB/OL]. [2023-10-21]. <https://arhiv.nuk.uni-lj.si/>.
- [32] BANQ. Archivage Web [EB/OL]. [2023-10-21]. https://www2.banq.qc.ca/collections/collections_patrimoniales/archives_web/?language_id=1.
- [33] Arquivo. pt. History [EB/OL]. [2023-10-21]. <https://sobre.arquivo.pt/en/about/press/history/>.
- [34] NetLab. Web Archive [EB/OL]. [2023-10-21]. <https://dighumlab.org/netlab/>.
- [35] Resaw. Resaw [EB/OL]. [2023-10-21]. <https://cc.au.dk/en/resaw/about>.
- [36] Internet Archive. About IA [EB/OL]. [2023-10-21]. <https://archive.org/about/>.
- [37] Internet Usage Worldwide. Statistics & Facts | Statista [EB/OL]. [2023-10-21]. <https://www.statista.com/topics/1145/internet-usage-worldwide/#topicOverview>.

苏依纹 中国人民大学信息资源管理学院博士研究生。研究方向: 档案信息化、数字人文。

周文泓 中国人民大学信息资源管理学院副教授, 硕士生导师, 博士。研究方向: 网络存档、档案与数据要素化。通讯作者。

刘鹏超 中国人民大学信息资源管理学院硕士研究生。研究方向: 档案信息化。