

档案领域词表自动化辅助构建 及知识组织应用探析

张昱 于薇

(中国科学技术信息研究所, 北京 100038)

摘要: 本文针对目前主题词表在实际应用中存在的词汇老化、结构陈乏, 以及无法适应当前知识服务等问题, 应用知识组织的方法对新型档案领域词表的构建进行研究, 并在此基础上, 从实际应用角度对档案知识组织体系构建及可开展的知识服务形式进行探究, 所提出的档案领域词表具备可扩展性, 其完善的知识体系、智能化的分析、个性化的服务能够有效提高档案服务的效率和能力, 更好地发挥电子档案的社会价值。

关键词: 词表构建; 知识组织; 知识服务; 档案

中图分类号: G350

DOI: 10.3772/j.issn.1673-2286.2018.06.010

21世纪以来, 信息资源的知识化管理和服务已经成为社会发展的一个重要标志。随着知识组织和知识链接技术的不断创新与突破, 信息与信息技术在国民经济和社会发展中扮演的角色越来越重要。在国家和社会全面信息化进程不断提速的大背景下, 知识组织与知识服务已成为现代档案工作的核心内容, 而构建档案领域词表是做好档案知识组织和知识服务体系的基础。

本文针对档案主题词表在实际应用中存在的词汇老化、结构陈乏, 以及无法适应当前知识服务等问题, 探讨基于知识组织的档案领域词表构建与应用, 从而提升电子档案的公共服务能力和水平。

1 档案数据知识组织研究现状

档案数据知识组织研究工作必须紧跟时代步伐, 面向应用、面向服务、面向社会, 以应用需求为导向, 不断升级技术体系, 完善词表结构与内容, 从而有效实现档案的知识组织与精准服务。贾玲等^[1]提出档案的知识组织是利用现代信息技术把档案知识源组织起来, 通过知识组织达到揭示档案知识资源内在联系, 并开展

知识服务的目标; 吕元智^[2]对数字档案资源知识组织的必要性和可行性进行深入分析, 提出数字档案资源知识组织框架, 建议加强档案资源管理领域的本体建设, 处理好档案知识元间的关联问题, 设计科学合理的知识关联获取模式; 李建忠^[3]提出档案信息资源知识组织方法包括分类主题法、主题地图法、元数据方法等, 强调档案信息资源的知识组织方法是传统档案信息载体知识组织方法的进一步深化, 更符合档案信息资源自身的特点和需要, 更有利于档案信息资源的利用和研究; 段荣婷^[4]以《中国档案主题词表》为例, 阐述知识组织与规范化控制的具体实现, 以及推进主题词表、分类法等知识组织系统的语义网络化应用。基于上述分析, 建议我国应尽快加强简单知识组织系统主题词表语义网络化的研究与应用, 从根本上实现一般性主题词表动态修订、维护、管理的电子可视化与网络化, 乃至语义网络化的扩展应用。

国际上, 在档案词表领域较有代表性的国家是荷兰和英国, 荷兰视听档案公共叙词表^[5]和英国档案叙词表^[6]已通过应用简单知识组织系统发布在语义网上, 极大地提高了包括档案在内的文化遗产的标引与检索利用效果。

2 档案领域词表构建

传统电子档案的知识组织主要基于《中国档案主题词表》和《中国档案分类法》。其中,《中国档案主题词表》是20世纪80年代国家档案局组织建设的,《中国档案分类法》于20世纪90年代由档案出版社出版。二者都存在建设时间久,后期没有及时更新维护的问题。如果采用人工审核方式为主、自动化方式为辅的方法对《中国档案主题词表》和《中国档案分类法》进行更新而形成新型的档案领域词表,将在一定程度上提高其应用的可行性。档案领域词表是在《中国档案主题词表》及其他相关领域词表基础上,基于可开放的档案领域语料,通过词表合并、新词发现、词间关系推荐、档案领域范畴的更新维护、词表逻辑一致性检测方法构建而成。

2.1 词表合并

词表合并旨在保留多个词表中同时出现的词及词间关系。合并操作需要先识别两个词表中词形相同的词条,根据词条相关信息(可能包含关系、属性等)的计算,给出词条全部合并的可信度,并进一步计算出两个词表合并的可信度,供用户处理参考。对于所有合并词条,都在界面上给出确认提示,经过加工者确认后,确定保留的词表部分。

2.1.1 词条合并可信度计算

输入:两个词条W1和W2的相关信息S1和S2。

$S1 = \{(P11, V11), (P12, V12) \cdots (P1m, V1m)\}$

$S2 = \{(P21, V21), (P22, V22) \cdots (P2n, V2n)\}$

输出:两个词条的可信度值如公式(1)。

$$T(W1, W2) = \frac{|S1 \cap S2|}{|S1 \cup S2|} \quad (1)$$

如“柱状晶”在有色金属(W1)、黑色金属(W2)词表中均出现。假设其在有色金属词表中有以下相关信息。

定义(P11):一种晶体形态(V11)。

属性-特点(P12):偏析比等轴晶少(V12)。

属性-特点(P13):结构致密(V13)。

属性-优点(P14):有较好的各向异性性能(V14)。

属性-特点(P15):具有抗蠕变能力(V15)。

分类-CLC(P16):材料(V16)。

分类-行业(P17):材料(V17)。

关系-铸锭(P18):层次(V18)。

则:

$S1 = \{(定义, “一种晶体形态”), (特点, “偏析比等轴晶少”), (特点, “结构致密”), (优点, “有较好的各向异性性能”), (特点, “具有抗蠕变能力”), (CLC, “材料”), (行业, “材料”), (铸锭, “层次”)\}$

假设词“柱状晶”在黑色金属词表中有以下相关信息。

定义(P11):一种晶体形态(V11)。

属性-特点(P12):偏析比等轴晶少(V12)。

属性-特点(P13):结构致密(V13)。

属性-优点(P14):有较好的各向异性性能(V14)。

属性-特点(P15):具有抗疲劳能力(V15)。

分类-CLC(P16):材料(V16)。

分类-行业(P17):材料(V17)。

则:

$S2 = \{(定义, “一种晶体形态”), (特点, “偏析比等轴晶少”), (特点, “结构致密”), (优点, “有较好的各向异性性能”), (特点, “具有抗疲劳能力”), (CLC, “材料”), (行业, “材料”)\}$

$S1 \cap S2 = \{(定义, “一种晶体形态”), (特点, “偏析比等轴晶少”), (特点, “结构致密”), (优点, “有较好的各向异性性能”), (CLC, “材料”), (行业, “材料”)\}$

故 $|S1 \cap S2| = 6$

$S1 \cup S2 = \{(定义, “一种晶体形态”), (特点, “偏析比等轴晶少”), (特点, “结构致密”), (优点, “有较好的各向异性性能”), (特点, “具有抗蠕变能力”), (特点, “具有抗疲劳能力”), (CLC, “材料”), (行业, “材料”), (铸锭, “层次”)\}$

故 $|S1 \cup S2| = 9$

两个词表中的词条“柱状晶”的合并可信度值为:

$$T(W1, W2) = \frac{|S1 \cap S2|}{|S1 \cup S2|} = \frac{6}{9} = 0.67$$

2.1.2 词表合并可信度计算

假设两个词表分别为Thesaurus1和Thesaurus2,二者有 k 对相同的词条,相应的相同词条的集合为 W ,则这两个词表合并的可信度 T 定义为公式(2)。

$$T = \frac{\sum_{w \in W'} (\text{Thesaurus1中的} w, \text{Thesaurus2中的} w)}{k} \quad (2)$$

这里,即两个词表合并的可信度为所有相同的词条的合并可信度的和除以相同的词条数目。

2.2 新词发现

采用基于启发式规则的、多测度融合的新词发现方法,通过线性加权,将词频、左右邻接熵、互信息结合进行判断,根据一定规则组合的候选词串是否可作为相应词表的新词,最终由领域专家审核。

(1) 词频。针对语料库中的所有候选词串,通过统计计算其频次,频次低于一定阈值的词串作为候选词串^[7]。

(2) 左右邻接熵。在自然语言处理领域,邻接熵被研究者广泛用来判定一个词串是否能构成一个合乎语法规则的词。通过信息熵来评测一个候选词与其左右邻接字符合并的可能性。

假设针对词串 t ,用字符 x 和字符 y 表示 t 的左邻接字符和右邻接字符,则本文采用以下公式来计算 t 的左邻接熵 $HL(t)$ 和右邻接熵 $HR(t)$ 。

$$HL(t) = -\sum_x p(x|t) \log p(x|t) \quad (3)$$

$$HR(t) = -\sum_y p(y|t) \log p(y|t) \quad (4)$$

其中, $p(x|t)$ 表示字符 x 是候选词 t 的左邻接字符的概率, $p(y|t)$ 表示字符 y 是候选词 t 的右邻接字符的概率。

(3) 互信息。互信息是新词发现中常用的统计量。假设对于候选词语“联合全宗”(词 t),如果想确定“联合”(词 x)和“全宗”(词 y)这两个字符在语料中的紧密程度,可按照公式(5)进行计算。

$$MI(t) = \log \left(\frac{p(t)}{p(x)p(y)} \right) = \log \left(\frac{Nn_t}{n_a n_b} \right) \quad (5)$$

值越大代表两者间的相关性越高,两个字符串连接后形成新词的可能性越高。通常可采用简单归一化

频率形式来估计概率。

$$p(t) = \frac{n_t}{N} \quad (6)$$

$$p(x) = \frac{n_x}{N} \quad (7)$$

$$p(y) = \frac{n_y}{N} \quad (8)$$

其中, n_t, n_x, n_y 分别表示 t, x, y 字符在语料中出现的频次, N 是集合中所有长度满足阈值(本项目设定为6)的候选字符串总数。

根据上述指标发现的候选新词,采用成词阈值筛选方法来确定最终的新词。具体方法如下。

由于用户提供的语料没有标注,因此采用无监督学习方法,即对词频、左邻接熵、右邻接熵和互信息4个统计量均设定一个阈值,如果一个候选字符串满足这4个阈值,就会被判定为一个新词。

$$\text{wordset} = \{t \mid \text{Freq}(t) > a_1 \wedge MI(t) > a_2 \wedge HL(t) > a_3 \wedge HR(t) > a_4\} \quad (9)$$

其中, a_1, a_2, a_3, a_4 分别表示词频、互信息、左邻接熵、右邻接熵的阈值。

基于以上计算方式,选用有价值的档案资料导入语料库,将达到一定阈值的词作为候选新词供专家审核,审核成功后列入词表。

2.3 词间关系推荐

在进行词间关系构建时,一方面根据汉语的组词特性,基于组词成分分析方法辅助构建词间等级、等同关系;另一方面采用深度学习技术构建词间相关关系。

(1) 等级关系。基于字面相似度算法,根据后方一致性原则进行词的入族处理并进行上下位类的划分,即根据字面相似度的结果,如果两个词包含相同的词素,且相同的词素位于词的后方,那么包含字数少的词作为包含字数多的词的上位词;反之,作为下位词处理^[8]。

(2) 等同关系。通过词汇间的前方一致、后方一致及两边一致3种途径进行字面匹配,识别同义词^[9]。

(3) 相关关系。针对给定的语料库和词库,根据word2vec工具,将词进行向量化,根据两个词间的动词向量均值或加权值等与系统中已有的关系进行对比,计算给定的词条与词库中的词条可能存在的关系。

2.4 档案领域范畴的更新维护

针对范畴表的更新,本文主要采用人工为主、机器为辅的更新方式。首先,邀请相关领域专家制订一级、二级类目;然后以不涉密的语料为基础,通过关键词查询到的文献覆盖度、主题挖掘方法确定三级及以下类目;最后,请领域专家进行审核,形成最终档案领域范畴表。主题挖掘主要基于LDA模型^[10],首先对文本语料进行分词、去停用词处理,然后调用主题模型(如LDA)挖掘主题,通过多种知识工程方法对LDA挖掘的主题补充标签。

2.5 词表逻辑一致性检测

对半自动化形成的档案领域词表进行逻辑错误检

测,可能出现的逻辑错误有:词间出现空格、全角等;族首词不能有属项;非正式主题词不可有属、分、代、参;一个词只能在其用代属分参项出现一次;词族中下位词的属、分不可跨层;代项不可为用项;两个词间如果具有上下位关系,不能再有参照关系;属分链中不能有非正式叙词;参项中不能有非正式叙词。通过设置程序可自动完成这部分内容的检测,借鉴论文^[11-12]中提到的基于图的检查方法,实现相应的一致性检查,发现错误并及时纠正。

3 档案领域词表的知识组织应用

在档案领域词表构建的基础上,本文从实际应用角度对档案知识组织体系的构建、可开展的知识服务形式等角度设计如图1所示的整体系统架构。

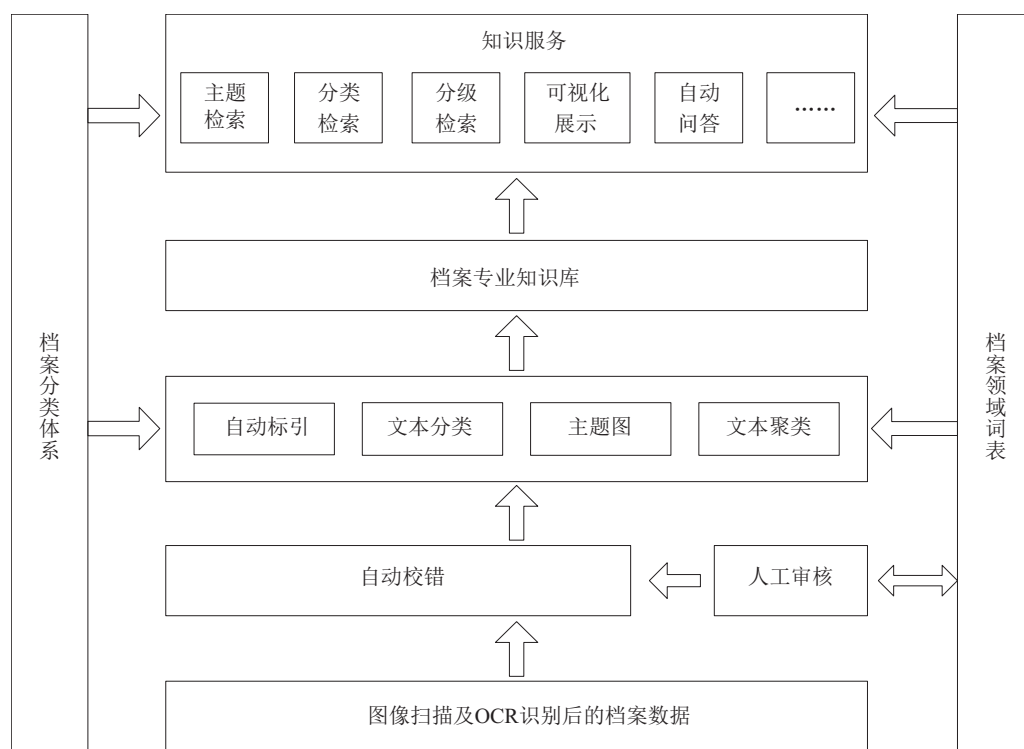


图1 整体系统架构

针对传统的纸质档案数据,首先采用图像扫描及OCR识别技术进行电子化,然后通过自动校错和人工审核的方式处理电子化的档案数据,保证数据的准确性。在更新后的《档案领域词表》支撑下,通过自动标引、文本分类、主题图和文本聚类等技术构建档案专业知识库,从而在档案专业知识库的基础上提供多种知识服务方式。

3.1 数据资源组织和检索

利用已构建完成的领域词表对电子档案数据资源进行标引,并基于标引结果提供基本检索服务。知识组织服务主要是知识资源的主题标引。主题标引以特定的领域词表为来源,精选代表性词汇进行标注,可进一步

以标注词汇代替资源进行简化计算。知识检索主要利用领域词表对用户的检索需求进行明确交互、对检索结果进行扩检和缩检,以及分类展示或聚类展示等。

3.2 知识导航

知识导航是利用已有领域词表的分类体系或上位概念对知识服务提供的科技文献、科学数据等资源,以及百科、图片等片段化条目知识进行组织,并在知识服务中按照分类体系进行树形关联导引,帮助用户逐步定位到所需资源和知识。

3.3 智能检索服务

(1) 针对知识问答提供知识的精准匹配。将重要问答知识与查询词条一一对应,针对查询内容给出标准的问答结果。

(2) 针对深度搜索提供检索结果的筛选与重新组织。基于通用检索策略得到的检索结果,利用已有领域词表对查询结果进行再次甄别,识别出不符合检索领域的结果,使检索结果更加精准,以便进一步用于情报分析。

(3) 针对知识地图提供领域知识概貌。利用领域词表的主要层次结构,提供对领域知识概览的支持;利用网状结构,提供对领域知识脉络展示的支持。

(4) 针对科技评价提供评价对象,界定评价范围。提供主要体现档案资料中关键事件、人物、机构的评价对象,便于确定科技评价中趋势分析等处理的对象范围,做好科技评价工作。

(5) 针对热点发现提供概念关联支撑。提供词汇间的关联,使得表现不同的词汇间能够通过概念层面建立联系,可以进一步提高热点分析的准确度。

4 结论

本文深度探讨档案领域词表的自动化辅助构建方

法、相应的知识服务架构、关键技术及可能的知识服务形式。构建的电子档案领域词表具有复用性和扩展性,可用于不同层次的电子档案数据挖掘的建设,基于领域词表提供电子档案的智能化分析、个性化服务将大幅提高公众获取档案的效率和能力,充分发挥电子档案的社会价值。

参考文献

- [1] 贾玲,刘要文,吕燕.论档案知识组织的方法[J].兰台世界,2012(14):31-32.
- [2] 吕元智.基于关联数据的数字档案资源知识组织研究[C].“新趋势、新思维、新途径”第六届“3+1”档案论坛”论文集,上海:上海世界图书出版公司,2012:17-25.
- [3] 李建忠.试论档案信息资源的知识组织与服务模式[J].档案管理,2013(1):49-50.
- [4] 段荣婷.基于简约知识组织系统的主题词表语义网络化研究——以《中国档案主题词表》为例[J].中国图书馆学报,2011,37(3):54-65.
- [5] VANASSEM M, MALAISÉ V, MILES A, et al. A method to convert thesauri to SKOS, 2006 [R/OL]. [2017-07-06]. https://link.springer.com/chapter/10.1007%2F11762256_10.
- [6] MILES A. UKAT (UK Archival Thesaurm) SKOS/RDF Data [EB/OL]. [2017-07-06]. <http://isegserv.itd.rl.ac.uk/skoa/ukat/>.
- [7] 杨阳,刘龙飞,魏现辉,等.基于词向量的情感新词发现方法[J].山东大学学报(理学版),2014,49(11):51-58.
- [8] 仲云云,侯汉清,杜慧平.电子政务主题词表自动构建研究[J].中国图书馆学报,2008,34(3):97-102.
- [9] 仲云云.电子政务主题词表的构建及应用研究[D].南京:南京农业大学,2007.
- [10] 唐晓波,房小可.基于文本聚类与LDA相融合的微博主题检索模型研究[J].情报理论与实践,2013,36(8):85-90.
- [11] 熊霞,常春,吴雯娜.叙词表相关关系逻辑检查方法的设计与实现[J].情报杂志,2010,29(11):154-158.
- [12] 徐硕,乔晓东,朱礼军.几种叙词表复杂逻辑错误检查算法研究[J].数字图书馆论坛,2010(8):55-58.

作者简介

张昱,男,1981年生,工程师,研究方向:科技信息管理、科技政策研究,E-mail: zhangy@istic.ac.cn。
于薇,女,1978年生,博士,副研究馆员,研究方向:科技档案、元数据技术。

The Construction and Knowledge Organization Application of Archives Thesaurus

ZHANG Yu YU Wei

(Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: In this paper the present situation of thesauri in the practical application of vocabulary, unable to adapt to the current issues of knowledge service, from the angle of knowledge organization construction of the new vocabulary in the field of archives building are studied, on this basis, the archives from the angle of practical application of knowledge organization system construction, and to explore the knowledge service form, the proposed archives knowledge organization system extensibility, the complete knowledge system, intelligent analysis, personalized service can effectively improve the efficiency of the archives service and the ability, more effective play to the social value of electronic archives.

Keywords: Vocabulary Building; Knowledge Organization; Knowledge Service; Archives

(收稿日期: 2018-04-13)

■ 书 讯 ■

《汉语主题词表》

《汉语主题词表》自1980年问世以后,经1991年进行自然科学版修订,在我国图书情报界发挥了应有作用,曾经获得国家科学技术进步二等奖。为适应网络环境下知识组织与数据处理的需要,由中国科学技术信息研究所主持,并联合全国图书情报界相关机构,自2009年开始进行重新编制工作,拟分为工程技术卷、自然科学卷、生命科学卷、社会科学卷四大部分逐步完成。目前工程技术卷和自然科学卷已出版。

《汉语主题词表(工程技术卷)》共收录优选词19.6万条,非优选词16.4万条,等同率0.84,在体系结构、词汇术语、词间关系等方面进行了改进创新。《汉语主题词表(自然科学卷)》共收录专业术语12.4万条,包含数学、物理学、化学、天文学、测绘学、地球物理学、大气科学、地质学、海洋学、自然地理学等学科领域,收词系统、完整,语义关系丰富、严谨,每条词汇都有相应的学科分类号表现其专业属性,并与同义英文术语对应。同时,建立《汉语主题词表》网络服务系统,提供术语查询、文本主题分析、知识树辅助构建等服务。《汉语主题词表》可用于汉语文本分词、主题标引、语义关联、学科分类、知识导航和数据挖掘,是文本信息处理及检索系统开发人员不可或缺的工具。

《汉语主题词表(工程技术卷)》已于2014年由科学技术文献出版社出版,分为13个分册,总定价3880元。

《汉语主题词表(自然科学卷)》已于2018年5月由科学技术文献出版社出版,分为5个分册,总定价1247元。两卷均可分册购买。