

数字图书馆隐私合规评估词典构建与应用^{*}

文禹衡^{1,2} 杨长元¹

(1. 湘潭大学知识产权学院, 湘潭 411105; 2. 湖南省数据治理与智慧司法研究中心, 湘潭 411105)

摘要: 按照选取种子词、形成语料库等步骤构建数字图书馆隐私合规评估词典, 以数字图书馆的隐私条款为基础构建语料库, 采用TF-IDF、TextRank、Synonyms等工具进行分词、词性标注和近义词拓展, 通过点互信息和共现分析方法分析内容词与种子词的映射关系, 实现合规点识别功能。经测试, 该隐私合规评估词典能够提高6.6%的隐私条款分词成词率、降低41.0%的冗余, 对合规评估点的识别准确率为90.5%, 在较高程度上实现评估点的自动识别, 提升数字图书馆隐私合规评估效率。研究有助于完善数字图书馆的通用用户隐私条款规范体系, 也为其他领域合规评估的智能化提供参考。

关键词: 数字图书馆; 词典构建; 隐私; 合规评估

中图分类号: G252 **DOI:** 10.3772/j.issn.1673-2286.2023.06.008

引文格式: 文禹衡, 杨长元. 数字图书馆隐私合规评估词典构建与应用[J]. 数字图书馆论坛, 2023 (6): 60-68.

2021年公布的《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》提出发展“智慧图书馆”的社会信息化战略。图书馆的设施资源、文献资源和人力资源迎来全面的智慧化、深度的智能化和服务的可及化^[1], 而这些离不开数字化和在线化。数字化是智慧化的前提, 数字图书馆是智慧图书馆的基础。在数字图书馆建设中只有重视用户隐私合规, 才能更稳健地推进智慧化图书馆建设进程。我国图书馆网站访问量在2019年达21.18亿人次^[2], 这意味着图书馆隐私合规将是不可忽视的议题。毕竟, “只有用户的隐私得到最大限度的保护, 图书馆也才能得到可持续发展的健康发展”^[3]。

图书馆服务中用户隐私保护的研究主要聚焦图书馆与读者之间的权利义务以及隐私条款的内容等方面。例如, 杨涛等^[4]提出图书馆用户隐私保护政策应当包含的内容和多途径公告的建议; 朱光等^[5]从收益、成本和损失的经济学视角出发, 构建多主体多策略的隐私

博弈模型, 并分析不同主体的博弈均衡策略; 周波^[6]分析数字图书馆个性化信息服务中的用户隐私问题, 从法律、技术、自律3个方面提出用户隐私保护策略; 易红等^[7]研究发现图书馆用户对隐私泄露容忍度并不高, 建议图书馆合法、合理、有节制地获取和使用用户个人信息数据; 马晓亭^[8]从法律、技术和数据分析3个方面研究图书馆用户个性化服务中潜藏的用户隐私风险, 提出相应的隐私保护对策与方法。

图书馆合规研究主要集中在数据安全、用户信息处理和隐私条款内容等方面。例如, 盛豪杰^[9]认为图书馆数据合规可以有效解决图书馆面临的数据安全内控缺欠问题, 提出图书馆数据合规的正当性、功能性和建议; 陈祥玲等^[10]梳理我国31个省市区图书馆官网隐私政策等文本, 提出图书馆处理用户信息的合规路径; 姜盼盼^[11]建议将图书馆隐私政策与国家立法作为我国图书馆隐私政策的合规依据, 合规内容包括隐私条款的形式可见性以及内容可读性。

收稿日期: 2023-05-03

^{*}本研究得到国家社会科学基金青年项目“数据要素确权的法律供给研究”(编号: 21CFX007)资助。

图书馆词典相关研究主要聚焦用户评论分词和检索等方面。例如,朱婷婷等^[12]借助移动图书馆评论语料、文献语料和相关词库构建移动图书馆词典,提高图书馆用户评论分词的准确率和时间性能;胡伶霞^[13]提出图书馆公共检索目录(Online Public Access Catalogue, OPAC)检索中基于词典的查询意图自动识别策略,该策略有助于检索系统更好地理解用户需求,提升检索结果与用户需求的相关度;王丽英等^[14]从数字资料库辅助工具的特性出发,以语文词典的编纂对古籍资料库的需求为解剖对象,提出服务于语文词典编纂的图书馆古籍数字资料库建设方向。

综上所述,尽管学界对图书馆用户隐私保护已有大量的研究成果,但是对于图书馆隐私条款(或称隐私声明、使用协议、法律声明等,以下统称为“隐私条款”)词典构建和合规自动评估的关注度还不高。在仅有的图书馆隐私条款合规研究中,学者仍通过人工方式展开合规评估,较少对图书馆隐私条款合规进行自动化评估,而与图书馆有关的词典仅涉及图书馆用户评论以及用户检索方面,还没有图书馆隐私词典方面的研究成果。鉴于此,本研究将对中国国家数字图书馆、省级图书馆和省会(首府)图书馆的隐私条款以及部分图书馆官方App(iOS端)的隐私条款进行调研,并以调研中收集的图书馆隐私条款为语料库,构建图书馆隐私条款词典,在此基础上实现对隐私条款合规评估点的自动检测,以促进我国数字图书馆用户隐私保护条款合规建设,推动建立规范的、通用的数字图书馆用户隐私条款。

1 词典构建方法

本研究旨在构建用于数字图书馆隐私条款自动合规评估的词典,其具有分词功能和隐私条款合规评估点自动识别功能。在实现合规评估功能前,词典只有分词功能,所以称“隐私词典”,在单独使用分词功能时也称“隐私词典”。基于研究成果与隐私条款现状,设计如图1所示的词典构建流程,主要可分为选定种子词、构建语料库、构建词典和实现评估功能4个步骤。

1.1 选定种子词

种子词是对合规状态下隐私条款结构化形式的概括,可以通过对结构的检验判断一份隐私条款在形式

上是否合规。因此,通过现有研究成果确定相对理想的隐私条款模型,并且借助专家从其结构中选取适量的词作为种子词。

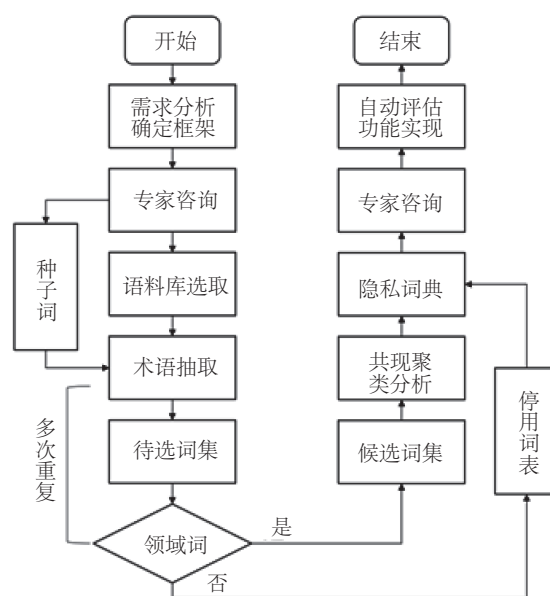


图1 数字图书馆隐私合规评估词典构建流程

1.2 构建语料库

目前,各个图书馆隐私条款的结构化形式不一,单纯通过种子词来进行判断过于片面,因此还需要选取各隐私条款内容中与合规评估点关联度高、有代表性的词作为“内容词”来进行补充。不能凭空编造内容词,故以中国国家数字图书馆、省级图书馆和省会(首府)图书馆的隐私条款和部分图书馆App的隐私条款为样本构建语料库,其能够反映图书馆隐私条款的现状,由此产生的内容词具有良好的代表性。

1.3 构建隐私词典

目前有3种构造分词词典的方法^[12]:①以人工输入词条信息为主、机器操作为辅;②从印刷版的词典里获取词条并手动录入;③对大规模文本采用简单的语言模型进行概率统计并分析有关词汇信息。本研究采用第3种方法,运用多种分词算法作为术语抽取工具,通过数据清理和人工的方式筛除不成词的结果和非领域词,构建隐私词典。

常用的术语自动抽取方法有基于字典和人工规则的方法^[15]、基于N-gram的方法^[16]、TF-IDF(Term

Frequency-Inverse Document Frequency)^[17]和TextRank^[18]等,其中TF-IDF和TextRank被广泛使用。TextRank可以更充分地利用文本元素之间的关系,更适合文本摘要任务,而TF-IDF更适合用于找出词频低但新鲜度高的关键词。结合TF-IDF与TextRank,以其他常用的分词方法为补充,运用Jieba工具^[19]进行术语抽取,再对分词结果进行人工审核,并使用近义词工具包Synonyms^[20]进行近义词拓展,从而构建隐私词典。

1.4 实现评估功能

自动评估功能需以隐私词典分词功能为基础,通过分词结果构建共现矩阵,计算词与词之间的点互信息(Pointwise Mutual Information, PMI)^[21],使用词共现分析方法分析词与合规评估点的关系。PMI是一种用来衡量两个具体事件关联强度的统计量,本研究用PMI来衡量词与词之间的关联度,关联度越高的词在文本中共同出现的频次越高,这两个词有极大可能描述同一个合规评估点。词共现分析方法是一种通过词共现信息进行量化分析的相关性分析方法^[22],其原理是将两个词记为一个词对,统计词对在同一篇文章中共同出现的频次,以此为基础对这些词进行分层聚类,从而发现词与词之间的亲疏关系^[23]。词共现分析方法可以基于PMI进行聚类,更好地分析出词与合规评估点的关系。

研究运用Gephi工具^[24]进行可视化分析,计算各节点词的平均加权重并使用模块化工具聚类。选取权重在平均值以上的词为内容词,分析其与合规评估点的映射关系。从分析结果中选取与合规评估点关联度较高的内容词,构建多对一的映射关系。通过判断分词结果中是否存在种子词和内容词,实现合规自动评估目的

(见图2),构建隐私合规评估词典。

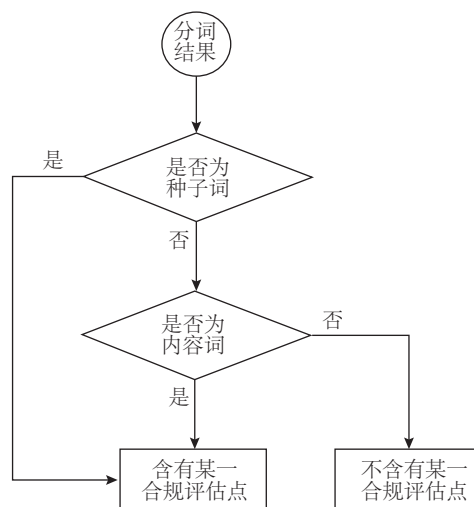


图2 数字图书馆隐私合规自动评估功能实现逻辑

2 词典构建与应用

2.1 词典种子词的选取

在《中华人民共和国个人信息保护法》的基础上,综合文献[10]、[11]、[25]和[26]的研究成果,构建隐私合规评估的基本框架,如图3所示,其中:适用范围相当于概述,告知用户该隐私条款的核心内容;用户信息处理包括图书馆网站对用户信息收集、使用和披露的各项规定,其中用户信息的收集还包括收集的目的、原则、方式和内容;权利包含用户的权利和图书馆的权利,用户权利包括查阅、更正、删除个人资料等,图书馆权利包括隐私条款的修改权和限制用户行为的权利等;用户信息安全包含用户信息存储方式和期限以及信息保护的技术手段和法律手段等;侵权救济是指当读

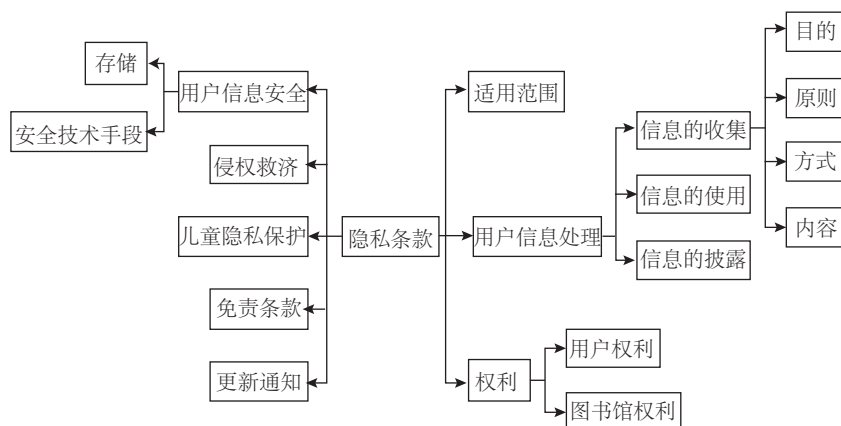


图3 数字图书馆隐私合规评估框架

者隐私遭到侵犯时,图书馆需要承担的责任;儿童隐私保护是指针对儿童这类特殊的用户群体提供的特别保护条款,如对儿童隐私的各项操作需要征得监护人同意等;免责条款是指图书馆需要明确在何种情况下,在发生泄露、丢失隐私信息等侵权情形时,可以免除自己的责任;更新通知是指图书馆在修改隐私条款的内容之后,应当通知用户注意条款内容的变化以及一些重大事项的通知等。

提出的框架是相对理想的隐私条款模型,从中选择各个末节点作为合规评估点,共形成15个合规评估点。将合规评估点的词适当拆分为种子词,以此构成种子词典,再通过术语抽取和人工筛选的方式确定领域词,对种子词典进行扩展,从而形成最终的隐私词典。种子词具体包括适用范围、收集、目的、原则、方式、内容、使用、披露、用户权利、图书馆权利、信息安全、存储、技术、侵权、救济、儿童、免责条款、更新和通知。

2.2 词典语料库的形成

用于构建隐私词典的语料库来源于图书馆隐私条款内容,在2022年9月20日至2022年10月20日,从中国国家数字图书馆和省级图书馆、省会(首府)图书馆的网站,以及App Store(iOS端)省、市级图书馆App中收集到8份隐私条款作为语料库的文本。之所以选择App Store而不选择Android应用市场收集图书馆App隐私条款,是因为前者的隐私保护要求较为严格。如此,尽可能进行全面调查以确保图书馆隐私词典语料库的全面性。

首先,访问中国国家数字图书馆和31个省级图书馆官网,主要在主页面、登录页面、注册页面和信息公开页面查找图书馆隐私条款。32个公共图书馆中,只有6个(18.75%)图书馆的官方网站显示隐私条款,分别为中国国家数字图书馆、福建省图书馆、广西壮族自治区图书馆、广东省图书馆、浙江省图书馆和内蒙古自治区图书馆。因江苏省图书馆官网无法访问,故视为不存在隐私条款;辽宁省和山西省图书馆网站虽然都设有“隐私声明”,但辽宁省图书馆官网“隐私声明”链接为空,山西省图书馆官网未在“隐私声明”中设置链接而不可见其内容;四川省图书馆数字馆藏门户要求用户通过微信小程序进行注册,而江西省和湖北省图书馆要求用户通过支付宝生活号进行注册,但无论是其官网还是第三方注册界面都没有显示隐私条款;剩余20个省级图书

馆网站均没有显示隐私条款。

在公布隐私条款的图书馆中,共获取3个样本。其中,中国国家数字图书馆、福建省图书馆、广东省图书馆和广西壮族自治区图书馆这4个图书馆的隐私条款的数量相同、内容相似,仅在“对读者资料实行保密”例外情形的举例上稍有差异,如福建省图书馆增加了一种情形“因通知等服务需要,需向短信运营商提供读者手机号等基本信息(若不同意,请不接受此协议)”,所以将这4个图书馆视为1个样本。浙江省图书馆的隐私条款就用户信息的收集、使用和保护作了规定,内蒙古自治区图书馆在用户个人信息的收集范围与方式、存储、获取、使用、共享、保护和管理方面都进行了说明。

其次,对27个省(自治区)的省会(首府)的图书馆进行调研,只查找到贵阳市图书馆官网存在隐私条款,计作1个样本。乌鲁木齐市和拉萨市尚未建设图书馆网站。因南京市、福州市和兰州市图书馆网站无法访问,故视为没有隐私条款。剩余21个图书馆均没有设置隐私条款。

最后,通过App Store进行搜索,查找到“首都图书馆”“上海图书馆”“吉林省图书馆”“山东移动数字图书馆”“掌上鄂图”“浙里阅”“浦东数字阅读”和“温州市图书馆”8款图书馆App,共获取4个样本。其中,“掌上鄂图”“山东移动数字图书馆”“浦东数字阅读”和“温州市图书馆”这4款App的隐私条款与成都超星数图信息技术有限公司开发的“超星移动图书馆”App使用的隐私条款相同,故视为1个样本。与该样本相比,“吉林省图书馆”App增加了“权限收集说明”,故视为新样本。“浙里阅”App的隐私条款与浙江省图书馆官方网站的隐私条款相同,不能视为新样本。此外,还有“首都图书馆”App和“上海图书馆”App 2个样本。

2.3 词典的构建

首先,将收集到的隐私条款文本进行预处理,使用txt格式存储,将文本中的英文字母全部转换为小写且删去其中全部数字和标点符号。其次,直接对话料库中每一份文本进行分词,将得到的分词结果与使用隐私词典得出的分词结果进行对比,检验隐私词典分词的有效性和效率。最后,使用多种方式进行分词。先使用TextRank算法加载种子词典,对话料库中的所有文本进行分词,共分出721个词,通过词性标注筛选出名词、

动词和动名词作为候选词,以人工方式审核分词结果,将结果中不成词的词语和非领域词添加到停用词表中。其中,需要对部分不成词的词或对领域词进行查阅修正,例如“机号码”和“安全流”实际上分别是被错误切分的“手机号码”和“数据安全流程”这类有意义的词。将修正后的词补充到种子词典中,使用TextRank算法再次进行分词,重复此过程,直到分词结果趋于稳定,最终确定225个词构成初始的隐私词典。

将上述隐私词典作为分词词典,使用TF-IDF算法进行分词,获得527个词,将这些词与现有词典中的词进行比对,经过人工审查筛选出未收录的29个词,将其加入隐私词典,累计得到254个词,另有273个词加入停用词表。

使用Synonyms分词工具和Jieba工具的精确模式分词重复上述步骤,发现隐私词典和停用词表内的词已经能够完全涵盖分词结果,证明基于语料库构建的隐私词典已经饱和,由此判断隐私词典构建基本完成,共得到258个词,其中名词204个、动词43个、动名词11个。通过Synonyms工具对现有词典中的词进行近义词扩充,再通过人工的方式对隐私词典进行近义词筛选,将近义词编入新的表格文件。在后续的分词中可以使用该近义词表,在读取数据后配合循环语句将分词结果与近义词进行匹配,匹配成功后用“现有词”进行替换,以便统一不同表达方式,对现有的停用词表去重后得到959个词。

2.4 自动评估功能实现

在隐私词典的基础上,通过聚类 and 人工审查方式对词典的功能进行拓展,旨在通过隐私词典对隐私条款完成分词后,直接得出该条款中存在的隐私合规评估点。基于PMI与词共现的原理,如若分词结果中出现

与种子词关联度高的词,其描述的内容很可能与种子词相关,表明隐私条款中可能存在该合规评估点,将这种词称为内容词。通过Pandas库^[27]读取包含“内容词-种子词”映射关系的Excel文件,将隐私词典分词结果与种子词进行匹配,即可实现对隐私条款的形式评估。但不同图书馆隐私条款的结构化形式各有不同,由种子词产生的结果缺乏一定准确性,故需要将分词结果与内容词进行匹配,对前一步操作进行补强,在一定程度上避免因隐私条款结构化形式不一而产生漏评。选取内容词以及构建“内容词-种子词”之间的映射关系,是实现自动评估功能的核心。

2.4.1 “词-种子词”之间的关联度验证

通过隐私词典进行分词,限制分词结果的词性为名词、动词和动名词,计算结果之间的PMI值,构建共现矩阵。实际取得194个词、1 522个词与词之间的关系(也称为“边”)。在计算每一条边的平均加权重后,通过Gephi的模块化工具——社区探测算法(Louvain)^[28],将解析度的值设置为标准解析度1.0,采用随机算法并使用边的权重作为考量参数,以便产生更好的分解和模块化结果^[29]。一般认为当模块度为0.3~0.7时,算法划分结果较为合适^[30]。最终的模块化结果将所有的词分为10个模块,编号为0~9,模块度为0.514,处于合理区间,各个模块的主要内容及词占比如表1所示。需要注意的是,模块词数量最少的模块9仅有两个词且词仅存在于App隐私条款中,占比仅为1.03%,故将其舍去。用户信息收集包含目的、原则、内容和方式4个方面,分置在模块0~2中,尽管模块9中也出现了“收集”,但仅限于“App权限收集”,与前述用户信息收集不同。

剩余9个模块的整体可视化结果显示,该词典主要与用户信息、数据、技术和保护4个词有较高的关联度:

表1 数字图书馆隐私条款分词各模块主要内容及词占比

模块编号	模块主要内容	词占比/%	模块编号	模块主要内容	词占比/%	模块编号	模块主要内容	词占比/%
0	用户信息收集原则、信息安全、儿童隐私保护	13.40	4	用户信息披露	5.15	8	用户权利、数据安全	4.12
1	用户信息收集内容	12.37	5	侵权救济、信息披露、免责条款	11.34	9	App权限收集	1.03
2	用户信息收集目的和方式	24.74	6	信息安全、免责条款	10.82			
3	信息披露、免责条款、使用目的	9.28	7	更新通知	7.73			

“用户信息”是隐私条款保护的主要内容;“数据”既可能是用户数据也可能指图书馆数据;“技术”可能代表收集用户信息的技术,也可能是保护用户隐私的技术;

“保护”是隐私条款的核心。但是，基于可视化结果难以确定具体的内容词以及与其关联度较高的种子词。

鉴于模块2词占比最大、模块8词占比最小，具体阐述模块2、8的可视化结果。模块2的可视化结果显示，该模块以数据、信息和技术3个词为核心，其中还包含种子词“收集”，可推测该模块与用户信息处理中“信息的收集”具有较高的关联度。除了上述词外，还有身份、IP地址、操作系统、浏览器、地理位置、网络日志、cookies和授权等与用户信息获取方式、信息来源相关的词，从而验证模块中的词与“信息的收集”具有一定的关联，其中还有小部分词（如统计、服务商和合作伙伴等）与用户信息收集的目的有关。因此，可以推断该模块的词与“信息的收集”有关联，具体涉及收集的目的、内容和方式，且该模块的词占比最大，可见语料库中的隐私条款比较重视用户信息的收集。

模块8的可视化结果显示其核心词包括身份验证、申诉、查阅和删除,由此可知模块8与用户权利相关。该

模块中的词都与身份验证一词相连,这是因为用户操作会引起数据变化,所以必须要验证用户身份,这体现对用户数据的保护。

通过模块可视化分析,发现语料库中的隐私条款样本比较重视用户信息收集、披露以及免责条款,与种子词相关联的词较为丰富,而在用户信息使用、用户权利、儿童隐私保护以及用户信息安全等方面还不够完善,与种子词相关联的词较少。

2.4.2 基于关联度的“内容词-种子词”映射取舍

基于分析结果,选取权重在平均值以上的边作为内容词的来源,得到183条边(366个词),经去重得到110个内容词。根据其与合规评估点种子词的关联度构建映射关系,因在映射中会出现多对多情况,故根据内容词与种子词权重进行取舍,优先选择权重较大的词,从而使“内容词-种子词”映射呈多对一的关系(见图4)。由图4可知,语料库中的隐私条款就信息收集内容和安全技术手段作了较细致的规定,而对儿童隐私保护、侵权救济和更新通知的规定就较为简略。

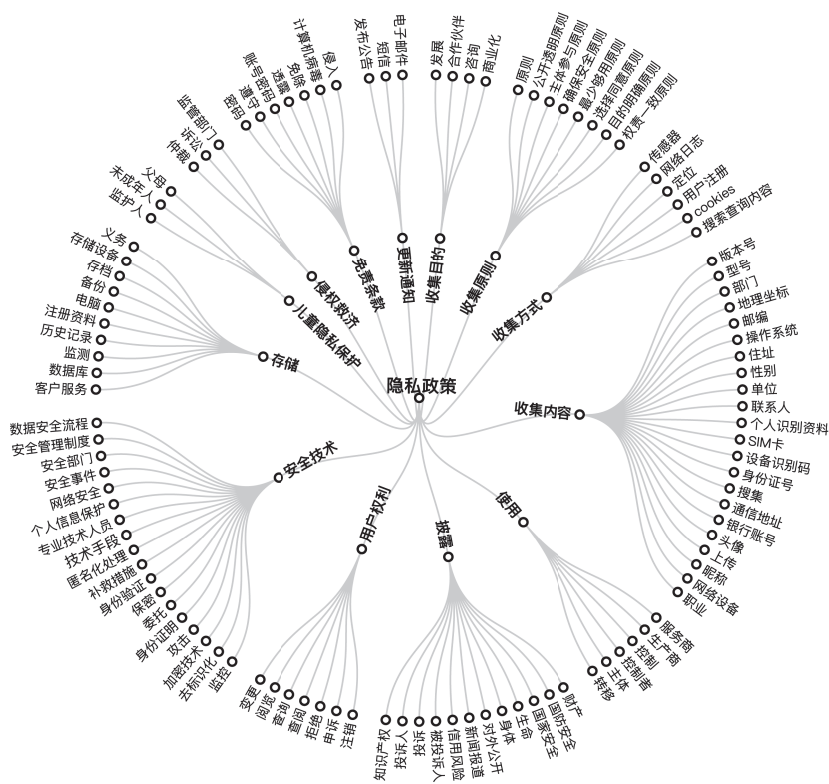


图4 隐私合规评估词典“内容词-种子词”多对一映射关系

3 结果验证与分析

3.1 词典分词效果检验

使用搭载隐私词典的分词工具依次对8份语料进行分词,与先前直接使用分词工具分词的结果进行比对:未搭载隐私词典的分词结果为1 372个,其中不成词结果99个,不成词率为7.22%;搭载隐私词典的分词结果为565个,其中不成词结果3个,不成词率为0.53%。从中可以得出,隐私词典能够有效降低不成词结果数量,减少分词结果的冗余,提高分词结果的正确率和有效率。对个人信息法律保护是一个持续且动态的过程^[31],隐私条款的内容也会随着相关法律的发展和现代科技

的更新迭代而产生相应的变化,所以该词典在今后的应用中也要作出相应的更新和调整。

3.2 自动评估效果检验

使用隐私合规评估词典对所有语料逐一进行测试,取得的隐私合规评估结果较好。为提升评估可信度,从高校及科研院所图书馆网站中随机选取若干样本用于测试,以验证该合规评估功能的可行性。

以“图书馆 隐私”为关键词在搜索引擎中进行搜索,选取中国科学院文献情报中心、国家哲学社会科学文献中心和湖南理工学院图书馆网站的隐私条款作为隐私合规评估词典测试的样本,测试结果如表2所示。

表2 图书馆隐私条款合规自动评估的测试情况

样本归属	合规评估点														
	适用范围	用户信息收集				用户信息使用	用户信息披露	权利		用户信息安全		儿童隐私保护	侵权救济	免责条款	更新通知
		目的	原则	方式	内容			用户权利	图书馆权利	存储	安全技术手段				
中国科学院文献情报中心	-	-	√	√	√	-	√	√	-	-	√	-	-	-	√
国家哲学社会科学文献中心	-	√	-	√	√	√	√	√	-	√	√	√	√	√	√
湖南理工学院图书馆	-	√	-	√	√	√	-	√	-	√	√	-	-	√	-

注:“√”表示自动合规评估的结果中存在该合规评估点;“-”表示其中不存在该合规评估点。

将隐私合规评估词典评估结果与样本进行对照之后发现:①中国科学院文献情报中心隐私条款包含用户信息收集的方式、用户信息收集的内容、用户信息披露、安全技术手段和更新通知,但因“原则”和“拒绝”这两个词,隐私合规评估词典误认条款中含有用户信息收集的原则以及用户权利两个评估点,通过提高“内容词-种子词”多对一映射质量或者解决“内容词-种子词”多对多映射选择问题能够有效减少此类误判;②国家哲学社会科学文献中心隐私条款包含用户信息收集的目的、内容、方式以及用户信息使用、用户信息披露、存储、安全技术手段、免责条款、更新通知、用户权利、儿童隐私保护、侵权救济,均被隐私合规评估词典识别;③湖南理工学院图书馆隐私条款包含用户信息收集的目的、方式、内容以及用户信息使用、用户信息披露、安全技术手段、存储、免责条款,均被隐私合规评估词典识别。

由测试结果可知:该隐私合规评估词典完全正确

识别出两个样本的合规评估点,针对样本的识别准确率为2/3(66.7%);中国科学院文献情报中心隐私条款的合规评估点数量为5个,词典识别出的数量为7个,其中包含5个正确的合规评估点,故该样本的合规评估点识别准确率为5/7(71.4%),其余两个样本的识别准确率都为100.0%,通过相加求平均值得到整个样本集的合规评估点识别准确率为90.5%。

3.3 检验结果分析

隐私词典可得到较好的分词结果,即隐私词典提高了约6.6%的成词率、降低了41.0%的冗余,有效提高了分词的效率,说明该词典分词的功能性较好,并验证了集合多种分词方法构造分词词典的可行性。

隐私合规评估词典实现了90.5%的合规评估点识别准确率,说明该词典的自动评估功能具有一定的精准度,验证了通过PMI与共现方法分析“内容词-种子词”

关联度, 以及在一定程度上使用内容词代替种子词代表合规评估点的可行性。同时, 也证明了通过该方法实现隐私条款合规自动评估的可行性。

4 结语

隐私条款是用户服务协议中不可或缺的基础条款, 是保障用户隐私安全的基本要求, 也是图书馆隐私合规的根本保证。当前学界对图书馆隐私条款的研究成果丰富, 但是并未聚焦图书馆隐私条款合规评估。究其原因, 可能是图书馆隐私条款语料数量不足, 尚未形成统一的合规评估标准。鉴于此, 本研究以中国国家数字图书馆、省级图书馆、省会(首府)图书馆网站和iOS端图书馆App的隐私条款为语料, 实现图书馆隐私合规评估词典的构建, 能够体现出已有隐私条款的“平均”水平。研究创新了隐私条款合规评估的方式, 通过构建隐私合规评估词典进行形式评估。在形式上, 后续可统一隐私条款的结构化形式并规范用词; 提高“内容词-种子词”多对一映射质量或者解决“内容词-种子词”多对多映射选择问题, 构建更加标准、科学的隐私合规评估词典; 将词典中的词与相关法条进行映射, 以提高合规评估的质量和说理性。以数字图书馆为示例, 使用计算机代替人工进行隐私合规评估, 有效提高了评估的效率和准确度, 推动隐私合规标准的统一, 保障隐私合规评估的客观性。其他领域也能够构建隐私合规评估词典, 以自主提高隐私合规程度。然而, 本研究还存在一些不足: 语料库质量不高, 映射关系选择可能具有一定的局限性, 且不同样本之间的形式结构存在差异, 例如用户信息披露中杂糅着免责方面的规定, 导致评测的结果存在误差。

参考文献

- [1] 王世伟. 论数字中国背景下的图书馆智慧化转型[J]. 中国图书馆学报, 2022, 48(1): 29-37.
- [2] 中华人民共和国文化和旅游部. 中国文化文物和旅游统计年鉴-2020[M]. 北京: 国家图书馆出版社, 2020: 56-65.
- [3] 梁荣贤. 基于“互联网+”视角的图书馆用户隐私保护研究[J]. 图书馆工作与研究, 2017(5): 124-128.
- [4] 杨涛, 曹树金, 冯彩芬. 图书馆个性化服务的用户隐私保护需求实证研究[J]. 图书情报工作, 2010, 54(5): 55-58.
- [5] 朱光, 曹雪莲, 于丹, 等. 图书馆情境服务环境下的多主体隐私博弈分析[J]. 情报科学, 2018, 36(6): 40-44, 72.
- [6] 周波. 数字图书馆个性化信息服务中的用户隐私保护[J]. 图书馆论坛, 2008, 28(5): 126-128, 21.
- [7] 易红, 任竞. 图书馆大数据服务环境下用户隐私泄露容忍度的实证研究[J]. 图书馆论坛, 2016, 36(4): 57-64.
- [8] 马晓亭. 大数据时代图书馆个性化服务读者隐私保护研究[J]. 图书馆论坛, 2014, 34(2): 84-89.
- [9] 盛豪杰. 图书馆数据合规: 缘起、原理及具体构建[J]. 图书馆, 2022(7): 1-9.
- [10] 陈祥玲, 肖冬梅, 杨忠. 图书馆处理个人信息的合规义务研究[J]. 图书情报工作, 2022, 66(17): 69-80.
- [11] 姜盼盼. 图书馆隐私政策合规性的依据与标准[J]. 图书馆建设, 2019(4): 79-86.
- [12] 朱婷婷, 郑德俊. 移动图书馆词典构建与应用[J]. 图书馆理论与实践, 2019(11): 75-78.
- [13] 胡伶俐. 图书馆OPAC检索中基于词典的查询意图自动识别[J]. 图书馆学研究, 2016(23): 72-76.
- [14] 王丽英, 王东海. 图书馆古籍数字资料库建设与质量评价新论: 以资料库建设与语词词典编纂的结合为例[J]. 图书馆理论与实践, 2011(2): 1-5, 27.
- [15] 王晓雪, 化柏林. 基于多源数据融合的公共文化领域词表构建研究[J]. 图书馆杂志, 2022, 41(10): 25-34, 96.
- [16] 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.
- [17] 徐文海, 温有奎. 一种基于TFIDF方法的中文关键词抽取算法[J]. 情报理论与实践, 2008, 31(2): 298-302.
- [18] 夏天. 词向量聚类加权TextRank的关键词抽取[J]. 数据分析与知识发现, 2017, 1(2): 28-34.
- [19] Joelost.Jieba[EB/OL]. [2022-10-10]. <https://github.com/joelost/jieba>.
- [20] WANG H L, HU Y X. Synonyms[EB/OL]. [2022-10-10]. <https://github.com/chatopera/Synonyms>.
- [21] 聂卉, 首欢容. 基于修正点互信息的特征级情感词极性自动研判[J]. 图书情报工作, 2020, 64(5): 114-123.
- [22] 阮光册, 夏磊. 基于词共现关系的检索结果知识关联研究[J]. 情报学报, 2017, 36(12): 1247-1254.
- [23] 路青, 靖彩玲, 范少萍. 基于互信息的共词分析方法研究[J]. 情报科学, 2016, 34(4): 48-51.
- [24] BASTIAN M, HEYMANN S, JACOMY M. Gephi: an open source software for exploring and manipulating networks[J]. Proceedings of the International AAAI Conference on Web and Social Media, 2009, 3(1): 361-362.

- [25] 罗曼. 论图书馆用户的隐私保护[J]. 大学图书馆学报, 2005, 23 (1): 63-65.
- [26] 张福学, 金标. 图书馆网络化服务中的用户隐私保护[J]. 图书馆理论与实践, 2001 (4): 17-18, 36.
- [27] Pandas-dev.pandas[EB/OL]. [2022-11-10]. <https://github.com/pandas-dev/pandas>.
- [28] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008 (10): P10008.
- [29] 邓君, 马晓君, 毕强. 社会网络分析工具Ucinet和Gephi的比较研究[J]. 情报理论与实践, 2014, 37 (8): 133-138.
- [30] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69 (2): 026113.
- [31] 文禹衡, 于琳. 我国个人信息法律保护现状、主要问题及完善路径: 基于《中华人民共和国个人信息保护法》的词频统计与分析[J]. 图书馆理论与实践, 2022 (4): 12-21, 28.

作者简介

文禹衡, 男, 博士, 副教授, 研究方向: 数据合规、数据隐私, E-mail: jsjmyh@163.com。
杨长元, 男, 硕士研究生, 研究方向: 数据合规、数据隐私。

Construction and Application of Privacy Compliance Evaluation Dictionary of Digital Library

WEN YuHeng^{1,2} YANG ChangYuan¹

(1. School of Intellectual Property, Xiangtan University, Xiangtan 411105, P. R. China;
2. Hunan Data Governance and Smart Justice Research Center, Xiangtan 411105, P. R. China)

Abstract: According to the steps of selecting seed words and forming a corpus, a privacy compliance evaluation dictionary of digital library is constructed. Based on the privacy clauses of digital library, a corpus is constructed. TF-IDF, TextRank, Synonyms, and other tools are used for word segmentation, part-of-speech tagging, and synonym expansion. The mapping relationship between content words and seed words is analyzed by pointwise mutual information and co-occurrence analysis methods to realize the function of compliance point identification. The test shows that the privacy compliance evaluation dictionary can improve the word formation rate of privacy clause segmentation by 6.6%, reduce the redundancy by 41.0%, and the recognition accuracy of compliance evaluation points is 90.5%, which realizes the automatic recognition of evaluation points to a higher extent and improves the efficiency of digital library privacy compliance evaluation. This study is helpful to promote the general user privacy clause standard system of digital library and provides references for the intelligent evaluation of compliance in other fields.

Keywords: Digital Library; Dictionary Construction; Privacy; Compliance Assessment

(责任编辑: 王玮)