中文个人名称规范记录的实体匹配与聚簇*

王瑞云　贾君枝

摘　要　本文尝试解决国内个人名称规范联合数据库检索结果集基于实体匹配的聚簇问题,分析国内名称规范联合库 CCCNA 的检索服务和数据库记录特点,提出对结果集记录合并聚簇的思路:首先预处理去除重复和明显的名称语义不匹配记录,再根据提取出的个人实体属性名称、出生年、个人关联的书目题名及关联的外部记录,基于个人实体的语义进行个人名称规范记录聚簇。实证统计结果显示,处理后结果集内的簇数都显著低于处理前的记录条数,与 VIAF 的关联聚簇结果也验证了本文方法的有效性。但本文书目匹配采取题名匹配,这会丢失一些有用的聚簇信息,后续研究将进一步集成图书机构的书目数据库,抽取更多的书目信息进行聚簇。图 4。表 6。参考文献 16。

关键词　虚拟国际规范文档　个人名称规范档　实体匹配　聚簇

Entity-Based Matching and Clustering of Chinese Personal Name Authority Records

Wang Ruiyun　Jia Junzhi

Abstract: This paper tries to deal with entity-based matching and clustering of retrieval result sets of Chinese personal name database. This paper analyses retrieval service of Cooperation Committee of Chinese Name Authory(CCCNA) and record features in the database and concludes that each retrieval result set has too many records needing to cluster based entity from semantics views. It first proposes preprocessing removing records of repeated and obvious mismatching between name and semantics, then proposes records clustering method based on names, birth-years, linked controlled numbers and books titles and links result clusters to VIAF. The results show that the quantity of clusters after processing is notably less than records before processing and the empirical study of linking into VIAF confirms the effectiveness of the methods. The book title is only taken as assistance of identifying of personal entity in the same person name recordings because of different reference formats and multilanguages from different references and we shall integrate bibliographic databases to present books based semantic entity in future research. 4 figs. 6 tabs. 16 refs.

Keywords: VIAF; Personal Name Authority Files; Entity Match; Clustering

1　引言

名称规范档根据控制规范规则,将名称信息按照统一的标目形式展现,为用户查询名称实体提供更高效的途径。由于单一机构构建的名称规范档规模较小,随着机构之间合作的深入,名称规范档资源之间的共建共享成为必然。2003 年,国家图书馆(NLC)、中国高等教育文献保障系统管理中心(CALIS)、香港地区大学图书馆协作咨询委员会(HKCAN)和台湾汉学研究中心(CCS)协同成立中文名称规范协作委员会(CCCNA),致力于中文名称规范档的共享使用。其提供的"中文名称规范一站式"查询系统,可实现对各种名称规范数据的查询,采用的模式是各家单位在物理空间上独立建库,通过网络环境来共享数据的分布式模型[1]。但 CCCNA 仅仅将各个机构之间查询结果合并显示,重合记录没有处理,单

* 本文系国家社科基金重点项目"基于关联数据的中文名称规范档语义描述及数据聚合研究"(项目编号:15ATQ004)的研究成果之一。

个库内及多库之间同一个人的记录没有聚簇,为用户的使用带来了极大的不便,也阻碍了与国外其它名称规范档的共享。

语义网技术的发展,进一步推动了跨机构的名称规范档、名称规范档与其它知识库的共享。现实世界由众多相互联系的实体组成。名称规范数据库中的记录是对实体的多方面属性的描述,将这些记录进行语义分析关联,则上升为语义网本体[2-4]。如何对中文人名规范档进行语义描述与关联成为研究重点[5,6]。其中与虚拟国际规范文档 VIAF 互联共享成为当前机构关注的问题,中文名称规范档与 VIAF 的互联共享将对中文名称的识别与归一具有重要作用,目前我国台湾地区已经处在测试阶段,旨在为其提供台湾地区已建设好的中文名称规范数据[7]。与 VIAF 连接过程中,如何有效地实现名称之间的匹配算法,成为连接共享的关键。Bennett 等学者开发了自动名称匹配算法,实现了德意志图书馆和美国国会图书馆规范记录的 70% 的自动匹配[8]。Hickey 给出处理 VIAF 规范文档的渐进精炼方法,先使用相当宽松的匹配形成粗糙集合,再将粗糙集改进为代表同一实体的所有记录的 VIAF 聚类[9]。

国内名称规范档与关联记录研究在图书情报领域已有一定的成果。贾君枝、白林林和李燕等人研究 MARC21 元数据与 CNMARC 元数据的比较与映射[10,11],张鹏图研究大英图书馆书目数据的关联化[12],贾君枝、石燕青研究中文名称规范档与 VIAF 的共享问题以及关联数据模型[1,13]。

基于上述分析,本文分析 CCCNA 的个人名称检索结果界面及详细页面的信息结构特征和统计特征,并实证批量下载 CCCNA 对 300 个中文个人名称检索结果的 MARCXML 格式文件,进行基于实体的语义研究。利用个人名称、出生年、关联外部记录及个人相关的书目,识别出描述同一实体的多条记录并进行聚簇合并,最后将聚簇结果进一步与 VIAF 关联,提高聚簇的准确度和簇的信息含量。

2 规范记录实体关联匹配分析

由于个人名称规范档中的人名普遍具有较高的知名度,本文从学科及相关的教材作者译者,诺贝尔奖获得者以及《中国人名大辞典(当代人物卷)》[14]选取 300 个中国现当代人名进行实验。根据人名中包含的汉字个数将其分为两大类:一类是两个汉字人名 100 个,另一类是三个及以上汉字人名 200 个。根据经验,假设两个汉字人名可能出现重名的几率大于三字及以上人名,每个名称检索出的结果集大于三个字及以上人名。

在 CCCNA 的“中文名称规范一站式”查询系统中[15],使用 300 个名称的简体形式进行检索输入,目标为检索到符合名称的所有个人名称记录。对于每次检索得到的结果集合,人工分析整个集合中各条记录与检索式的匹配情况(真实匹配还是伪匹配)、重名记录出现的情况等。希望从检索结果分析中明确当前中文名称规范记录四大库中各个库内部记录的特点,探寻名称实体识别合并过程中可依据的信息,为下一阶段自动化实体聚合的探索做准备。

2.1 检索结果及重名记录分布

由于存在重名的人物,更重要是 CCCNA 不精确的匹配算法,导致检索命中的结果集过大。表 1 分别从匹配记录、检索记录数进行统计,可以看出 NLC、CALIS 两者记录命中率较高。同样,其检索两个汉字的结果中出现的重名甚至伪匹配记录(汉字名称和注释串接的字符串包含检索名称)也最多,伪匹配超过准确匹配的记录数。检索结果存在数据库内部聚合和跨数据库的同一实体聚合的可能。

2.2 各个库内部记录的特点

NLC 的规范记录库中,存在记录号不同、规范标目及内容完全相同的记录且占比很高,本实验中有 178 个人名出现此类问题,由于这类记录不是完全重复记录,因此记录归并难度增加。

CALIS 将中文简体、中文繁体、西文、日文等同时作为并列规范标目,一定程度上丰富了规范标目的形式,为读者提供了更加便利的检索途径,

但是具有相同记录控制号的同一条记录在检索结果中重复出现多次,结果集变大,干扰项或不符合项太多,降低了查准率。CALIS 记录中的参考数据源包含书目题名信息,同时建立了与百度百科、维基百科的链接,可以将人名记录实体扩展关联到这些知识库,进一步丰富实体识别判断依据,可进行更广泛的实体关联。

HKCAN 将中文名称作为连接标目,等同于其他数据库的规范标目。其所提供的数据来源较广泛,其中书目名称中有大量英文题名,因此如果与中文书目题名匹配涉及到多语言处理功能;同时其多数记录中有一个外部链接记录号,链接到美国国会图书馆,对与国外名称规范的联接具有重要作用。

CCS 命中记录较少,明显低于其他三个本地库,有大量的参考数据源字段,可以用来挖掘书目题名,但也存在多语言匹配问题。但 CCS 已处在 VIAF 的测试阶段,对名称聚簇到 VIAF 的关联有重要参考价值。

表 1　CCCNA 各个库的检索覆盖和匹配记录统计

记录数 \ 数据库名	NLC	CALIS	CCS	HKCAN
两字名称检索有匹配记录的集数(总 100)	97	92	68	95
三字及以上名称检索有匹配记录的集数(总 200)	185	166	94	135
两字名称检索记录总数	1140	1691	599	381
三字及以上名称检索记录总数	372	629	139	153
两字名称单个检索结果集最大记录数	160	230	142	48
三字及以上名称单个检索结果集最大记录数	12	199	16	5

2.3　匹配检索点的构成及特征

能够进行匹配聚合的检索点取自检索结果中的字段和数据库记录内的字段,各备选检索点统计信息如表 2 所示。CCCNA 的规范标目(HKCAN 是连接标目)包括中文简体名称、生卒年,这些可看作是最强检索点信息。关联外部记录号可以通过关联的外部记录准确地关联该组记录到 VIAF 聚簇,是强检索点。规范记录中的附注和参考数据源中的出版翻译书目,对个人实体有较强的辨识作用,可作为强检索点。其它信息为弱检索点信息,包括名称附注、变异名称、附注、参考数据源。

CCCNA 四个来源库中,HKCAN、CCS 的参考数据源结构较复杂,有多种情形:除强检索点的个人出版翻译作品外,还有描述研究个人的传记作品,收录个人和众多其他人的各种形式的名人录、作家笔名录,而最后一种情况与个人实体的关联性弱,不能用来匹配。

表 2　CCCNA 中备选检索点

检索点名称	包含记录数	检索点特征
中文简体名称	5104	最强匹配点:构造粗糙的原始检索结果集,全部记录均包含,但伪匹配和重名多。
生卒年	2636	最强匹配点:各机构统一,适于跨机构、多语言聚簇;但若干记录该信息缺失。
书目题名	4035	强匹配点:包含在附注、参考数据源中,语义上对实体匹配作用强;但存在多语言题名困难,跨机构书目信息格式复杂,匹配准确度受限,有些书目与人名弱关联。
外部链接记录号	486	强匹配点:关联外部记录,丰富记录信息,供与 VIAF 等关联使用。但包含该信息的记录少。
记录控制号	5104	弱匹配点:用于单个库内细分记录,不能供实体聚合,但是记录标识关联的键。

续表

检索点名称	包含记录数	检索点特征
名称附注	2304	弱匹配点:其中有 1486 条出现在 CALIS 中,分类体系多,包括职业、专业、民族、性别等,缺乏分类叙词表支持,匹配作用弱。
变异名称	3601	弱匹配点:包含该信息的记录较多,有简繁体、多语言形式,辅助中文简体名称匹配。
附注	2177	弱匹配点:大段文本,仅挖掘书目题名。
参考数据源	5061	弱匹配点:多种机构多格式书目信息,供挖掘出版翻译书目题名。

3 自动聚簇实现

CCCNA 系统提供在线检索和 APP 下载导出数据,在线检索结果页面如图 1 所示,记录对应的详细页面如图 2 所示。APP 下载提供了 4 种输出格式,包括文本格式、MARC-2709、MARC-文本格式和 MARCXML 格式。本文选择第 4 种输出格式 MARCXML 格式作为实验的源数据。

MARCXML 文档包含了在线检索的结果和详细记录信息,CCCNA 利用工具对其中规范标目和其他内容进行语义转换处理,形成结构化文档,如表 3 所示。文档外层采用 XML 语法,包含便于机器处理的元语言标记,用 record 标签标识结果集中每条记录的起点和终点。记录内层字段采用 MARC 标记,语义上与 MARC21 和国内图书馆业界规范 CNMARC 相近,参考国内学者对 MARC21 格式的解析以及 CNMARC 与 MARC21 格式的映射[6,16],我们可以分析出 XML 记录内部字段的语义。每个检索结果集以 MARCXML 格式下载到本地保存,作为实验的原始输入数据,分别存放在本地以检索名称编号的 300 个 XML 文件中。

图 1 CCCNA 查询结果网页

图 2 图 1 第 5 条记录详细内容网页

表3　查询结果网页、输出XML文档与数据库表字段的对应关系(＊表示没有该部分信息)

检索结果		输出XML文档记录		RDF映射		数据库表字段	数据库表字段含义说明
结果总页面	记录详情页面	XML记录字段	子字段				
＊	＊	＊	＊	＊		SNO	检索人名编号
＊	＊	＊	＊	＊		Sname	检索人名
记录来源	记录来源	801	$a	schema. Orginization		SDBC4	来源库
＊	记录控制号	001	$a	＊		SRNO	记录控制号
＊	外部控制号	009	$a	＊		SRLO	外部记录控制号
标目值	中文个人规范名称	200(第1条)	$a	schema. person	schema. name	CName	中文简体规范名称
			$f		schema. birthdate schema. deathdate	BDY	生卒年
			$c		schema. description	CN2	人名附注
＊	西文个人规范名称	200(第2条)	$a		schema. name	EName	外文规范名称
＊	变异标目名称	400(多条)	$a		schema. name	VNames	变异名称集
＊	标目信息附注	300	$a		schema. description	Note	附注
＊	参考数据源	810(多条)	$a	schema. book	schema. name	Refdatas	参考数据源集合(以‖分隔)
＊	＊	＊	＊	schema. book	schema. name	books	书名集合,由Note和Refdatas计算得到

3.1　数据库表的构建

每个XML文件包含检索结果集中的所有记录(包含重复记录),建立字段与数据库表属性对应表(见表3)。编写程序算法从每个XML文件分析出所有的记录,从每条记录中提取结构化属性:检索名称、记录控制号、来源库、中文简体名称、生卒年、名称附注、英文名称、变异名称、备注、参考数据来源等,运行程序计算得到书目题名集合,形成两个数据库表。

算法:从XML文件构建数据库表

输入:100个两个汉字检索结果全部下载的XML文件;200个三个及以上汉字全部下载的XML文档,每个文档根据检索名称的编号命名,如amark2002.xml。

(1)构建数据库表的结构,设计表的属性,共21个,预先设计了计算属性和分簇属性。

(2)从XML文档中根据标记record标签划分出每条记录,对应用户检索结果页面的一行及其链接,对应输出数据库表中的一行。

(3)循环处理XML划分出每一条记录,记录中字段与属性的映射关系与含义如表3所示。

1)按照表3的对应关系处理XML的当前记录,数据库表插入一行记录,各属性为XML文档映射的值;

2)进入下一条记录的文本位置,处理下一条记录,直到文档结束。

输出:两个数据库表,表的属性字段相同,处理方法也相同,所以采用一个算法。

3.2　数据预处理

数据预处理是对两个数据库表中的记录使用SQL语句过滤全部属性完全相同的记录,根据名称语义处理删除不能正确匹配的伪匹配记录。表4显示预处理前后两个数据库表的记录数,可以看出预处理使结果集中的记录数大大减少,提高了用户检索的便利性。

预处理前后,各检索结果集中的详细记录数的累积概率分布情况如图3所示。为了显示清晰,左右图形中只给出累计覆盖主要的90%结果

集的记录分布,对于检索名称结果集中记录条数异常多值的没有显示出来,两个表中两步预处理后记录数的累积分布曲线都在处理前曲线的左上方。尤其左图两个字的预处理后曲线的改进效果更加显著,两步都比处理前显著向左上方移动。

表4　两个初始数据库表及预处理前后的基础统计信息

记录数	两个汉字名称检索	三个及以上汉字名称检索
检索名称总数	100	200
初始下载得到(预处理前)的记录总数	3811	1293
删除相同记录后剩下的记录数	2959	983
删除名称伪匹配后剩下的记录数	1016	879
初始单个检索结果集最大记录数	580	204
删除相同记录后剩下单集最大记录数	391	104
删除名称伪匹配后剩下单集最大记录数	90	35

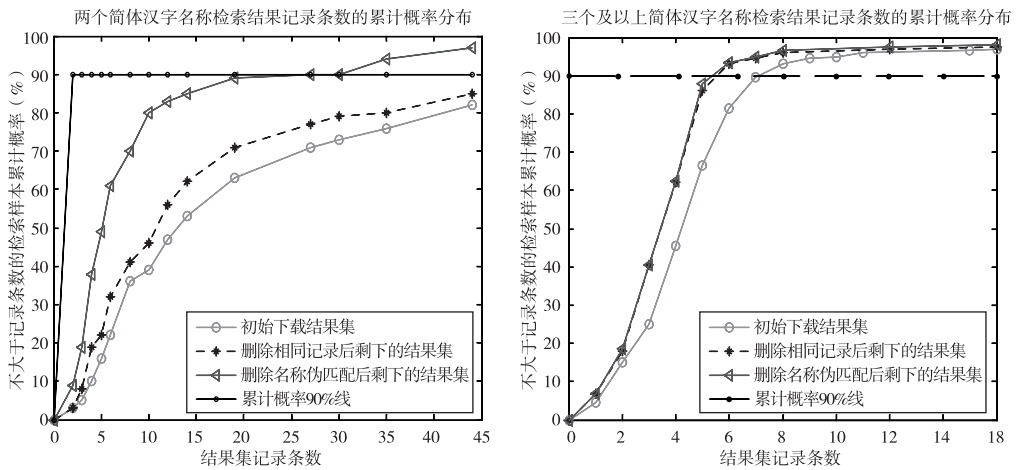

图3　两个、三个及以上汉字名称检索结果集预处理前后记录数累积概率分布

3.3　名称实体匹配和聚簇算法

将同一个实体的记录划分到同一个聚簇记录,合并多条记录信息,具体实现思路如下:

(1) 将个人名称相同的记录归类

将名称相同的记录归入一个记录集,共300个记录集。利用中英文名称、变异名称信息,合并该集合中与检索名称语义相同的记录。

(2) 对每个记录集中的记录按照出生年聚簇

先按出生年排序再聚簇。由于生卒年份属性有两部分信息,由"~"分割的字符串,进行分开计算,得到出生年和卒年,排序和聚簇要按照单个标准,又因为本实验数据的名称实体取自当代或现代名人录,没有卒年信息的很多,所以聚簇算法选择按出生年排序后再聚簇。排序后的记录,出生年字段不为空的,进行聚簇,同一出生年的记录分为一簇,分配同一个簇号,不同的出生年划分到不同的组,按顺序分配不同簇号,检索名称编号和组号一起形成聚簇号;对于出生年字段缺失为空的记录,由于信息缺失不能聚簇,每一条记录划分为单个记录的一簇,分配不同的簇号,一条记录分配一个聚簇号。

(3)总体聚簇

总体聚簇对确定性的提高体现在出生年空缺的记录。两簇中有相同记录号的记录合并两个聚簇为一簇。优先向有出生年的聚簇关联合并;两簇中的记录,书目题名相同的合并为一簇。书

目题名的语义相同,需要进行语义计算:

先对每条记录在附注中根据“《》”符号提取出每个书名,每个书名在书目字段中 Books 用|| 分割串联,再对参考数据源中的作者译者类型条目提取书目题名,串接在 Books 字段后。每条记录的 Books 都是个人书目的不完全集合,只包含本数据库收集的书目。

对分属于两簇的两个书目集合,两个集合中的每一项分别比较匹配,一个集合中的任一项匹配上另一集合的一项,两个就合并为一簇,两个书目集合做并集运算成一个集合。由于多语言问题和参考数据源的不同性质类型,只能根据书目题名文本匹配辅助判断两本书可能是相同内容作品,从而推断同名的两簇集合的作者为同一个人的可能性超过合并的阈值,进行人名记录簇的合并。书目题名语义匹配还需要借助多语言问题研究成果,应用不太成熟,本文将在以后研究[16]。

总体聚簇还要处理出生年相同的簇是否确实是同一实体,是否需要分裂为两簇。出生年相同的簇内记录,查找额外的匹配支持信息,只要满足下面任意一条就不再分裂:(1)规范名称和变异名称合起来有两个以上相同;(2)记录号相同;(3)书目题名语义相同;(4)名称附注相同。如果所有上述匹配支持信息都没有,总体聚簇只好分裂,划分出一个新簇。

实体匹配聚簇算法输出:对记录集内进行实体匹配聚簇后,分配得到出生年聚簇号,总体聚簇号。

3.4 聚簇结果分析

根据上述聚簇算法,运行得到两阶段的聚簇结果数据(表 5)。可以看到两种聚簇结果从总簇数、单结果集最大簇数都体现出明显的合并效果,簇数都远远小于算法运行前的记录数。一次检索结果合并到一个实体簇的覆盖率比聚簇前都有显著提高,从 5%提高到 45%,原因是检索结果的多条记录实为同一个人的信息。除了合并到一个簇的极端情况,图 4 中两类聚簇后的两条累积曲线都显著地位于聚簇前曲线的左上方。这说明不管是否重名或重名多少,每个结果集的记录都得到合并,结果集簇数缩小,单个簇是合并记录的并集,信息得到较大丰富。

表 5 聚簇算法运行前后基础统计信息

类 型	聚簇前	出生年聚簇(总)	总体聚簇
300 次检索簇数	1895	1101	1137
有出生年的总簇数	1434	640	738
无出生年的总簇数	461	461	451
有外部记录号的总簇数	197	196	195
单个名称的最大簇数	90(2072)	53(2072)	79(2072)
单个名称的最大簇数(有出生年)	66(2072)	29(2072)	50(2072)
单个名称的最大簇数(无出生年)	34(2032)	34(2032)	34(2032)
90%的检索集最大簇数	10	7	8
合并为一簇检索集占全部检索集(300 个)的百分比(%)	5	43	45

注:聚簇前为记录数,聚簇后为簇数。

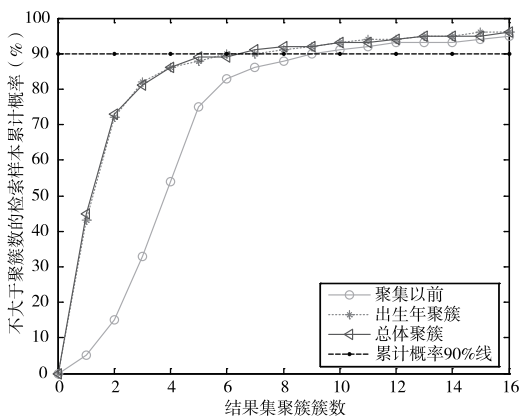

图 4 聚簇算法前后结果集中聚簇数的累积概率分布

3.5 聚簇结果与 VIAF 匹配聚簇验证

利用聚簇算法的结果,将 300 个结果集中的 1137 簇与 VIAF 匹配聚簇,与聚簇的记录数 5104 条记录比较,检索匹配次数大大减少,每次匹配的检索点信息又很丰富。根据本文结果的聚簇及其语义与 VIAF 匹配聚簇:首先根据结果簇中的外部记录控制号和 CCS 的记录控制号与 VIAF 匹配关联,再在剩下的结果簇中分有无出生年的簇与 VIAF 匹配聚簇,具体处理结果如表 6 所示,是一个比较可行且有效的到 VIAF 的匹配聚簇方法。

表6 结果聚簇与VIAF的匹配聚簇

处理工作	拟匹配簇数	剩余簇数	匹配结果	处理建议
外部控制号或CCS的记录号与VIAF簇匹配	407	730	准确匹配	同簇内其他记录加入匹配的VIAF簇
剩余簇中有出生年的簇	426	304	较准确匹配	名称生卒年匹配簇内的记录加入匹配的VIAF簇,匹配不上增加一个高质量的新簇。
剩余簇中无出生年的簇	304	0	不精确匹配	增加新簇到VIAF,新簇质量不高。

4 研究局限与展望

本文对个人名称聚簇通过实体属性姓名、生卒年以及关联书目题名基本上建立了个人名称的实体匹配和记录关联,局限是书目题名只是根据题名的文本信息,只能在相同姓名的记录集内辅助支持个人实体的识别,达不到对书目的实体识别,不利于建立广泛的个人名称和书目名称的实体关联,书目实体匹配关联需要进一步获取更多的书目数据支持。

论文下一步的研究方向有两方面:一,借助多语言信息组织和检索研究[16],考虑不同语言数据库的书目题名实体匹配问题;二,参照本体或叙词表的研究成果,对附注信息和参考数据源进行语义标注,得到关于名称更丰富的属性信息,以便更加准确高效地进行个人名称和书目题名的实体匹配和关联,进行更精确的知识关联。

参考文献

1 贾君枝,石燕青.中文名称规范文档与虚拟国际规范文档的共享问题研究[J].中国图书馆学报,2014(6).

2 陆伟,武川.实体链接研究综述[J].情报学报,2015(1).

3 贾君枝,等.汉语框架网络本体研究[M].北京:科学出版社,2012:7-16.

4 曾建勋.知识链接及其服务研究[M].北京:科学技术文献出版社,2012:16-18.

5 曹宁,仲岩.论中国个人名称标目的区分问题[J].中国图书馆学报,2006(6).

6 郝嘉树,王广平.中文人名规范的语义描述与关联探讨[J].图书情报工作,2012(14).

7 VIAF: The Virtual International Authority File [EB/OL]. [2016-07-20]. http://www.oclc.org/research/project/VIAF.

8 顾犇.虚拟国际规范文档——连接德意志图书馆和美国国会图书馆的规范文档[J].国家图书馆学刊,2006(4).

9 T B Hickey, J A Toves. Managing Ambiguity in VIAF[EB/OL]. [2016-09-14]. http://www.dlib.org/dlib/july14/hickey/07hickey.html.

10 贾君枝,白林林.关联数据中CNMARC到MARC21的映射的实现[J].国家图书馆学刊,2015(4).

11 李燕,等.MARC21元数据与CNMARC元数据的比较分析[EB/OL]. [2016-08-13]. http://www.nlc.gov.cn/old2008/service/fuwudaohang/conf2006/conf2006_liyan.htm/.

12 张鹏图.大英图书馆书目数据的关联化分析[J].国家图书馆学刊,2015(4).

13 贾君枝,石燕青.中国个人名称规范文档的关联数据化研究[J].情报学报,2016(7).

14 廖盖隆,等.中国人名大辞典——当代人物卷[M].上海:上海辞书出版社,1992:1-23.

15 中文名称规范协作委员会.中文名称规范联合数据库检索系统[EB/OL]. [2016-09-22]. http://cnass.cccna.org/jsp/resultlist.jsp.

16 司莉,贾欢.2004—2014年我国多语言信息组织和检索研究进展与启示[J].情报学报,2015(6).

(王瑞云 讲师 山西大学经济与管理学院管理科学与工程专业2007级博士研究生, 贾君枝 教授 山西大学经济与管理学院)

收稿日期:2016-12-09