馆藏资源语义化关键技术及实证研究*

楼 雯

摘 要 本文从微观层面设计了馆藏资源语义化模型,描述了馆藏资源语义化的关键技术,并利用武汉大学图书馆馆藏资源"美洲各国军事"类目的数据对模型进行了检验。从馆藏资源到语义资源需要经过信息提取技术、语义关系提取技术和形式化技术的支持。实验分析发现馆藏资源语义化模型所述的流程可用,后续研究可以着眼于资源统一化。图5。表3。参考文献53。

关键词 馆藏资源 语义化 分词 概念提取 关系提取 形式化

分类号 G250

An Empirical Study on Key Technologies of Library Resource Semantization

Lou Wen

ABSTRACT This paper designs a model so as to figure out the entire process and key technologies of library resource semantization. The model was tested with America's military category data from Wuhan University Library's bibliographic retrieval systems. Interchange from library resources to semantic resources are supported by key technologies such as information extraction, relationship extraction, and formalization. Empirical analysis supports the feasibility of the model and further research should be focused on resource consolidation. 5 figs. 3 tabs. 53 refs.

KEY WORDS Library resources. Semantization. Segmentation. Concept extraction. Relationship extraction. Formalization.

语义网的提出至今已有十几年时间,人们对语义网环境下的生活充满期待,众多机构和个人将身边的信息资源发布成语义信息,使之成为语义网的一部分。信息资源语义化的形式有很多种,凡是将人们掌握的知识通过先进技术转化成机器能够理解的语言,都可认为信息被语义化了,所以发布语义信息的途径不仅仅是构建成本体或关联数据。但本体和关联数据是目前学者们首肯的语义化方式,近年来,世界著名机构如 BBC[1]、路透社[2]、维基百科[3]、美国国会图书馆[4]、中国国家科技图书文献中心[5]等,纷纷将其资源语义化,在互联网上发布和提供查询。2007 年,W3C 设计了全民关联数据的计划,最大程度地接近了语义网。

信息资源语义化已经成为知识交流和知识共享的必经之路,图书馆作为蕴含巨大信息资源和知识的集合,馆藏资源的语义化在世界一些地区已经成为语义网建设的重要组成部分,在另一些地区也即将成为重点研究的对象。语义网经过十几年的时间还未能实现,不仅仅是浩瀚的信息海洋造成的,也因为语义化过程中会遇到种种逻辑难题和技

* 本文系国家社科基金重大项目"基于语义的馆藏资源深度聚合与可视化展示研究"(批准号:11&ZD152)的研究成果之一。

通讯作者:楼雯,Email: hotwen_l@sina.com

术难题。语义网的实现是一个层层推进的过程,首先将一部分易于语义化的现有资源语义化,进而带动其他部分的语义化,而图书馆就是现成的实验对象。本文专门为馆藏资源的语义化设计了模型,总结归纳了语义化过程的关键技术,并用两个实验揭示了模型和关键技术的可行性,旨在为馆藏资源的语义化进程提供参考。

1 相关研究

馆藏资源首先是一种信息资源,是所有信息的一部分,其次是图书馆特有的资源,再次馆藏资源经过人类和机器的理解转化为知识,所以它还是一种显性的知识资源。馆藏资源具有这三个含义,其语义化的相关研究也可以从这三个方面进行分析。

(1)语义网的关键技术

实现语义网的技术是连接馆藏资源语义化与万维语义网的关键,目前的研究多以总结语义网技术和提出新型语义化技术为主。在本体研究的热潮中,相关学者已将语义网的关键技术默认为本体及其相关技术,这一类的研究包括了全面介绍语义网信息组织的技术和方法,并总结出 RDF、OWL 和本体是语义网的核心技术[6-7]。本文强调广义的语义化,因此这些总结出的关键技术并不能代表所有的语义网技术。在新型技术的研究上,有学者提出了语义化网络的学习算法[8]、知识的自动分类技术[9]、微格式技术[10]可以作为语义网实现的关键技术,但这些技术的使用环境较为局限,研究也缺乏全面性。当然也有学者认识到总结归纳语义网关键技术的必要性[11],但只描述了问题,并未解决问题。

(2)数字图书馆的关键技术

数字图书馆在馆藏资源从数字化到语义化的过程中起着重要作用,数字图书馆相关技术的研究包括了对语义化技术的应用以及微观、中观层面技术的研究。在技术的应用方面,目前学者偏向于利用 RDF、元数据、本体和关联数据[12-15]进行图书书目的语义化或提出新的知识组织方法,也就是说这些学者将 RDF、元数据、本体和关联数据视为数字图书馆实践中的关键技术。另外,微观层面的技术包括了概念提取、概念转换[16-18]、互操作、语义互联、概念格[19-23],中观层面的技术包括了语义网格、SOA[24]、本体构建、本体映射、本体进化[25-27],可以看到,这些研究重点描述了数字图书馆语义化的某种技术,并没有形成一套完整的流程和技术体系。

(3)知识服务的关键技术

语义网和数字图书馆的建设和实现实际上都是为了知识交流和知识共享,因此上文提到的许多研究已经表现出知识组织或个性化服务的内容和关键技术[9,15-16,19-22]。不仅如此,有的研究总结了聚类技术、数据挖掘技术在图书馆建设中的具体使用方法[28-30];认为手工决策技术、基于内容的推荐系统、基于本体的服务系统和智能信息推拉技术是个性化服务的技术支持[31];提出基于读者行为的知识服务关键技术有读者特征提取技术、兴趣模型分析技术和协同推荐技术[32]。这些是对知识服务技术特征的总结和探讨。还有文献[33]利用关联数据将多种数据源的知识关联到一起形成语义扩展,则是对关键技术的应用。

近年来,随着馆藏资源语义化进程的加快,一些学者提出了有建设性的语义化模型和框架,为馆藏资源语义化和知识服务提供了参考[34-36]。上述研究有的仅设计了模型或进行实验验证,有的则仅描述一部分技术,尚缺乏对馆藏资源语义化过程整套系统关键技术的归纳总结。因此本文设计了馆藏资源语义化模型,并描述其中各个部分的关键技术。

2 馆藏资源语义化模型

馆藏资源语义化主要包括信息的提取、语义关系的提取、形式化和应用等步骤,图 1 显示了馆藏资源语义化的主要过程。

数字时代的图书馆已不再是纸质图书的集合地,现有的馆藏资源有很多种,若要把所有馆藏资源语义化,则要考虑到所有形式的馆藏资源。

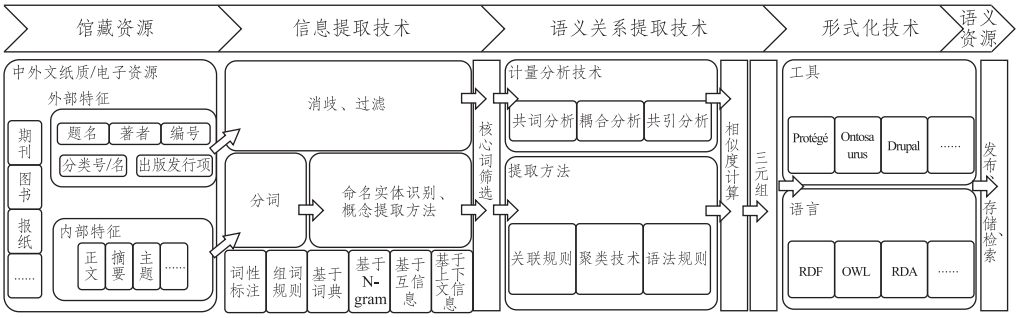

图1 馆藏资源语义化模型

尽管不同的资源类型有不同的行文格式和出版样式，但是它们均具有外部特征和内部特征。外部特征包括题名、著者、编号、分类号/名和出版发行项等，内部特征则包括正文、摘要和主题等。馆藏资源语义化的过程中，内外部特征的语义化内容不同，语义化的过程也不同，因此需要区别对待。

对内外部特征分别预处理后，则进入信息的提取步骤，这一步主要运用的关键技术统称为信息提取技术。内部特征是表示资源主题内容的信息集合[37]，表示方式多为文字段落，需要进行分词处理才能将内部特征显现出来，在分词过程中，由于语言语种的不同，分词的方法和注意事项也不同，总体来说分词时需要考虑分词算法、词性标注和组词规则。经过分词后的段落已经是零散的信息点，要提取有用的信息点，运用到的关键技术是概念提取技术，概念提取技术又可以进行细分。外部特征的概念提取没有内部特征繁杂，外部特征是已经被主题标引后的信息，可直接视为概念，但仍需进行消歧、过滤等处理，防止重名、特殊情况的出现。

内外部特征经过核心词筛选后，就可以进行语义关系提取的步骤。核心词主要依靠已有叙词表和领域专家来筛选。语义关系提取技术包括计量分析技术和提取方法。外部特征之间的关系可以很清楚地用计量分析技术表现出来，这里所说的计量分析可以是文献计量分析、信息计量分析、科学计量分析和网络计量分析，分析方法有共词分析、耦合分析、共引分析等。而内部特征之间的关系种类繁多，需要利用文本中词之间的关系体现出来，词之间有等级和非等级的关系，可以通过基于关联规则的、基于聚类的、基于语法规则的提取方法得来。

不论内部特征还是外部特征，概念和关系的强弱都需要经过相似度计算才能确定。当概念和关系都提取出来后，就可以形成三元组，对三元组形式化，从而将馆藏资源转换成语义资源。

3 馆藏资源语义化关键技术

3.1 信息提取相关技术

信息提取是指从结构化信息、半结构化信息和非结构化信息中提取概念或实例并将其存储成事实信息的过程[38]。结构化信息和半结构化信息（如文献外部特征）提取概念较为方便，从非结构化信息（如文本）提取概念需要对文本中的字词进行取舍，如何判断取舍则需要分词技术、命名实体识别和概念提取技术。

（1）分词技术

分词技术是自然语言处理的研究范畴，国内外学者对自然语言理解展开了深入的研究。西文分词方法大致可归为三大类：基于语法的分析法、基于语法与语义相结合的分析法和基于语义的分析法三类，常用的分词工具有 Lucene、SimpleAnalyzer、WhitespaceAnalyzer 等。汉语分词方法有基于词典的分词方法、基于统计的分词方法、基于理解的分词方法[39]。基于词典的分词方法需要一个标准词典，一般用正向最大匹配算法、逆向最大匹配算法

和最小切分算法使待分词文本与词典匹配,匹配成功的词则被切分;基于统计的分词方法不需要词典,如果把词看作固定的字的组合,相邻的字共同出现的次数超过一定阈值,则把这些字当作一个词;基于理解的分词方法是机器学习的算法,机器在分词的同时进行语法、句法和语义分析,需要经过大量的学习试验集才能确定精度。

词性标注就是利用计算机给文本中的词标上词类,如"电脑"是名词、"美丽"是形容词等。词性标注有助于机器识别。自然语言常会出现词组、兼词(一个词具有多个词性)和新词,给词性标注带来很大困难,组词规则可以解决一部分难题。虽然有现有的词性词典和组词规则,但使用时还要考虑到实际情况,有些专有性强的分词文本更多利用专用叙词表。

国内自动分词系统主要有清华 SEG 分词系统、复旦分词系统、北大计算机研究所分词系统和中国科学院 ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System)[40-41]。ICTCLAS 主要功能包括中文分词、词性标注、命名实体识别、新词识别,系统基于层叠隐马模型(Cascaded Hidden Markov Model, CHMM)而设计,利用了基于词典的和基于统计的分词方法,具有开源的特点,应用较为广泛。

(2)命名实体识别

命名实体识别是信息处理技术的关键基础技术,命名实体是文本信息中的基本单位,是固有名称、缩写等的唯一标识[42]。命名实体识别即发现命名实体并进行类型的标注。命名实体的识别可分为命名实体的识别和新词的抽取两种类型。通用的识别过程一般依据《中国人名大词典》、《世界地名翻译大辞典》等现有的资料与文本进行匹配,并标注上实体类型[43]。除词典形式的已有资料外,国内外学者已建立了通用本体的新的组织方式,更快捷准确地识别命名实体。正是在现有资料的基础上,像 ICTCLAS 均可以实现命名实体识别的功能。但基于词典的识别方法有很大的局限性,作为新词的命名实体无法被标注,这时则需要利用概念提取方法。

(3)概念提取方法

分词后的文本已经成为概念的离散的集合,有些可能是错误的概念,需要概念提取方法将其完善。总结多年来学者的研究,提取语义概念的方法有基于词典的方法、基于 N-gram 的方法、基于互信息的方法、基于上下文信息的方法和混合方法[44],其中基于互信息的和基于上下文信息的方法有助于提取合成词。①基于词典的方法,又称为基于规则的方法,该方法有一套标准的词典与分词后的结果进行匹配,匹配成功的词则成为待选概念。这种方法提取出的概念精准度高,但方法的约束性太强,符合标准的自然语言或相似词均无法被提取出来。②基于 N-gram 的方法,这种方法将相邻的 N 个分词文本中的词组合起来,形成新概念,如"人们/n 的/u 脑子/n 里/f 就/d 会/v 出现/v 英勇/a 的/u 形象/n"的 2-gram 结果为"人们的"、"的脑子"、"脑子里"、"里就"、"就会"、"会出现"、"出现英勇"、"英勇的"、"的形象",可以看出结果往往出现不是词的概念,错误率较高。因此有时根据实际情况会多次选择不同的 N 来提高准确率。③基于互信息的方法,互信息是统计语言学模型中度量两个词之间关联程度的指标,通过计算 A 词和 B 词相邻出现在文本总词数中的概率,确定合成词 AB 是否为概念词。④基于上下文信息的方法,概念之间的上下文与概念有紧密联系,设 A、B 词的上文同有 C 词,计算此情况出现的概率,概率越高说明合成词 AB 越可能是概念词。⑤混合方法就是将以上方法取长补短,搭配使用后提取概念的效果更佳,比如可以先用基于词典的方法将标准词提取出来,再将剩下的通过 N-gram 算法形成新词,利用基于互信息的或基于上下文的方法对新词进行过滤,最终形成整套的待选概念。

3.2 语义关系提取相关技术

语义关系多被分为等级关系和非等级关系,提取方法大为不同。等级关系是树型结构,与聚类技术的结果类似,因此多用层次聚类算法提取;非等级关系在现实生活中出现更多,形式多样,比如地理位置关系、人物关系、属性关系等,因此提取方法也有多

种,如关联规则、计量分析方法和语法规则等。

(1)相似度计算

相似度计算贯穿概念提取和关系提取的始终。常用的计算方法有四类:①基于特征的计算方法,两个概念若拥有的共性越多,说明两者相似度越大,反之则差异性越大,这种方法又被称为 Tversky 指数[45],其可变形为 Tanimoto 系数、Dice 系数等。②基于距离的计算方法,其基本思想是计算出的两个概念的距离越远,则相似度越低,反之则相似度越高,在本体中一般利用概念距离根节点的路径长度计算两者距离。③基于信息论的计算方法,两个概念拥有的共同信息越多,说明相似度越高[46],信息论的方法是基于特征的计算方法的变形,共有信息的度量只能依靠共有特征的度量。④混合方法是通过概念的同义词集、语义邻居概念和概念特征多重指标综合计算概念间的相似度[47]。

(2)聚类技术

聚类技术是数据挖掘的重要方法,不同的聚类方法可以提取到不同概念之间的关系。国内外研究聚类技术依对象可分为一维聚类、二维聚类和多维聚类,或概念聚类、词聚类、文本聚类和文献聚类;依算法可分为基于划分的方法、层次聚类方法、基于密度的方法、基于网格的方法和基于模型的方法。①基于划分的方法将聚类对象划分为几个初始组,初始组被反复迭代进行优化直至不能再改进,划分法的著名算法有 K-means 算法及其变形[48],Clarance 算法[49]等;②层次聚类方法是一种自底向上或自顶向下的方法,先将聚类对象独立成每个原子类,再利用某些相似度规则进行逐层聚类,主要算法是 CURE 算法[50];③基于密度的方法不同于层次聚类方法中计算原子类之间的距离,而是从聚类对象的密度出发,主要用于空间数据的聚类,典型算法是 DBSCAN 算法[51];④基于网格的方法将数据划分为有限的单元格,再分析单元格中的数据进行聚类,如 STING 算法[52];⑤基于模型的方法是将某一聚类对象抽象成一个模型,再从其他的对象中寻找最优的和模型匹配。

(3)计量分析技术

计量分析能够确定概念之间的关系强度,如果说本体在表示概念及其相互关系时,偏重如何将其表现得更有内涵,而计量分析则能够帮助本体挖掘出概念间是何种关系[36]。计量分析的主要方法有共词分析、耦合分析和共引分析等,提供的计算结果是两个概念共同出现的次数,从而确定两者的关系强度。举例来说,作者 A 和作者 B、作者 A 和作者 C 共同撰写过某些文章,定义两者关系均为“合作”,这是浅层次的关系,B 和 C 究竟谁与 A 合作密切,计量分析的结果可以表达出来,从而得到“强合作”关系和“弱合作”关系等,也可以用具体的相似度数值表现强弱关系。

值得一提的是,关联规则也是数据挖掘的重要方法,利用关联规则可以发现概念之间潜在的语义关系。关联规则的原理是从原始数据集中找出高频集,利用高频集产生规则,也可用规则检验数据项是否满足条件,经典的算法有 Apriori 算法[53]。语法规则也是提取非等级关系的关键,非等级关系中谓词的选择主要依靠语法规则进行提取,西文和汉语都有特定的语法将字和词组成句子,如一个完整的句子至少包括主谓宾结构,分析语法结构可以分析句法,从而了解词和句子的语义关系。

3.3 形式化技术

无论是哪种语义资源,都需要承载工具才能将其发布到语义网中,这种承载的工具就是形式化语言和工具。由 W3C 领衔的语义语言开发已经形成了规模,Metadata、RIL、WSDL、RRL、XTM、RDF、RDFS、XML、OWL、FOAF、DC、RDA 等语言的出现,丰富了形式化语言的内容。形式化工具依托形式化语言被开发出来,如适用于 RDF、RDFS、XML 和 OWL 的 Protégé、Jena、Apollo、Ontolingua 和 WebODE 等工具,关联数据发布工具 D2RServer、Drupal、DBpedia 等。

4 实验与讨论

为了验证馆藏资源语义化模型的可用性,体现语义化关键技术的可行性,下文分别针对馆藏资源的内外部特征设计了实验,揭示馆藏资源语义化过

程。实验数据为随机选取的武汉大学图书馆馆藏书目检索系统中的 E7(中国图书馆分类号)“美洲各国军事”类图书信息。

4.1 实验一——馆藏资源外部特征的语义化过程

(1)在武汉大学图书馆中,E7 类图书共 167 种,将其按全记录格式下载并预处理后,得到 307 个著者(含团体著者),75 个出版单位, 179 个标引主题词,将武汉大学图书馆书目标引卡片格式的各个字段作为语义概念,则得到如图 2 所示的类目体系。

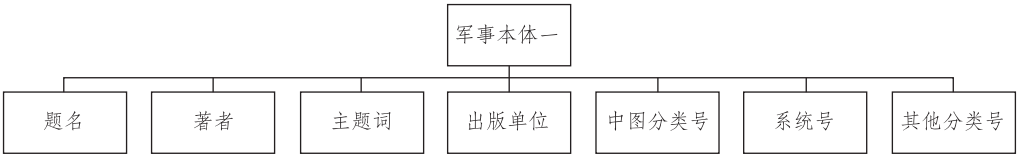

图 2 军事本体一类目等级体系

(2)定义类目属性,包括类目本身的属性(Object Property)和类目的数据属性(Data Property),类目本身的属性表示各类目之间的关系,比如题名类隶属于军事本体一类,题名类由出版单位类出版等等;数据属性规定了类目数据的特征,比如其他分类号的数据类型为双精度型,系统号不能为空等。类目之间的关系与出版事实相符,不再赘述。

(3)添加实例。将著者、出版单位等概念设置为各类目的实例,此过程中需要确定实例的准确性,即进行核心词汇的筛选,因为题名、分类号、系统号和出版单位均为固定数据,只需检查即可,所以筛选的主要内容是著者和主题词。

(4)定义实例属性,即提取实例间的关系。类目体系中的一些实例关系是特定的,比如题名被分到某一分类号,出版单位出版某一本书,某著者撰写某一题名,这是浅层的语义关系。另有一些实例的关系较为复杂,比如著者之间的关系,主题词之间的关系,著者和主题词之间的关系,有著者合作关系,主题词共现关系,著者主题词共现关系等。下面仅举例说明著者合作关系的提取,307 位著者中有合作关系的作者 215 位,两两合作次数为 400,所以共有 400 对著者合作关系,在整个合作网络中不同关系对的关系强弱不同,利用共现分析得到两两著者的合作次数,并按基于特征的相似度计算公式 $S(c_1,c_2)=\frac{2\times f(c_1\cap c_2)}{\left[2\times f(c_1\cap c_2)+f(c_1-c_2)+f(c_2-c_1)\right]}$ 得到两两著者的共现关系强度(见表 1)。其中,$f(c_i)$ 为著者单独出现的次数,$f(c_1\cap c_2)$ 表示 c_1 和 c_2 共同出现的次数,$f(c_1-c_2)$ 表示 c_1 出现而 c_2 不出现的次数,$f(c_2-c_1)$ 表示 c_2 出现而 c_1 不出现的次数。在表 1 中,语义相似度只有 2 个数值,但整个合作网络肯定会出现更多的数值,在进行语义标注时不能将数值作为关系表示,于是要将语义相似度抽象化表示,比如将相似度值为 0.67 的关系表示为高度相关,0.4 的关系表示为中度相关,则得到如表 1 所示的著者语义关系对及三元组。

(5)形式化。利用 RDF 语言将所有三元组形式化,最后形成军事本体一,如图 3 所示。

4.2 实验二——馆藏资源内部特征的语义化过程

(1)武汉大学图书馆馆藏书目检索系统中有关图书内部特征的标引字段有内容简介、摘要、网络摘要,本文选取 E7 类图书的内部特征作为原始数据,167 本图书中只有 99 本有此内部特征,将这些内部特征下载并存储成 TXT 文档。

(2)分词。本文利用中国科学院的分词软件 ICTCLAS,将图书的文本内容进行切分,分词的结果如表 2 所示。

(3)概念提取及核心词汇的筛选。将所有文本合并为一个文本,利用基于 N-gram 的提取方法,结合组词规则,将文本分词结果提取成概念词。利用《英汉军事大词典》进行核心词汇的筛选,从而

表 1　军事本体一中著者两两关系语义相似度计算过程($f(c_1)\geqslant 3$)

c_1	c_2	$f(c_1)$	$f(c_2)$	$f(c_1\cap c_2)$	$f(c_1-c_2)$	$f(c_2-c_1)$	$S(c_1,c_2)$	相关性	三元组
莫里斯	蔡晓惠	3	3	2	1	1	0.67	高度相关	<莫里斯,高度相关,蔡晓惠>
莫里斯	符金宇	3	1	1	2	1	0.4	中度相关	<莫里斯,中度相关,符金宇>
莫里斯	靳绮雯	3	2	2	1	1	0.67	高度相关	<莫里斯,高度相关,靳绮雯>
莫里斯	林贤明	3	1	1	2	1	0.4	中度相关	<莫里斯,中度相关,林贤明>
蔡晓惠	靳绮雯	3	2	2	1	1	0.67	高度相关	<蔡晓惠,高度相关,靳绮雯>
蔡晓惠	林贤明	3	1	1	2	1	0.4	中度相关	<蔡晓惠,中度相关,林贤明>
蔡晓惠	米琳	3	1	1	2	1	0.4	中度相关	<蔡晓惠,中度相关,米琳>
蔡晓惠	莫里斯	3	3	2	1	1	0.67	高度相关	<蔡晓惠,高度相关,莫里斯>
蔡晓惠	墨菲	3	1	1	2	1	0.4	中度相关	<蔡晓惠,中度相关,墨菲>
赫恩	白堃	3	1	1	2	1	0.4	中度相关	<赫恩,中度相关,白堃>
赫恩	胡升新	3	1	1	2	1	0.4	中度相关	<赫恩,中度相关,胡升新>
赫恩	李进	3	1	1	2	1	0.4	中度相关	<赫恩,中度相关,李进>
赫恩	易亮	3	1	1	2	1	0.4	中度相关	<赫恩,中度相关,易亮>
赫恩	郑金艳	3	1	1	2	1	0.4	中度相关	<赫恩,中度相关,郑金艳>

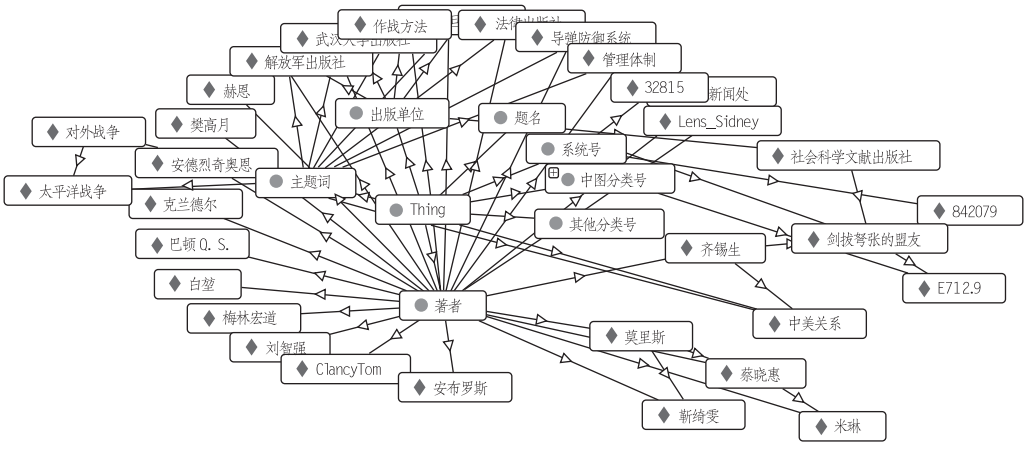

图 3　军事本体一(部分)

组成本体的类目体系(见图 4)。

(4)语义关系的提取。首先确定概念之间的等级关系,利用层次聚类的算法将相似度较高的一批概念对提取出来,比如“国家”和“半殖民地国家”、“冷战时代”和“后冷战时代”、“关系”和“合作关系”等。再利用基于距离的相似度计算方法计算出邻近词的相似度,提取高相似度的邻近词,比如“同盟关系”和“合作关系”、“互动行为”和“关系”、“组织体制”和“体制革新”等。

第二步确定概念之间的非等级关系,概念之

表 2　图书文本分词结果(部分)

只要/c 一/m 提起/v 巴/j 顿/q ,/w 人们/n 的/u 脑子/n 里/f 就/d 会/v 出现/v 一个/m 英勇/an ,/w 威严/an ,/w 暴躁/a ,/w 善战/v 的/u 美军/n 司令官/n 形象/n 。/w 人们/n 称/v 他/r 为/v “/w 有/v 指挥/v 大军/n 的/u 天才/n ”/w ,/w 特别/d 擅长/v 进攻/vn ,/w 追击/v 和/c 装甲/b 作战/v 。/w《/w 巴顿/nr 将军/nz 战争/n 回忆录/n 》/w 是/v 他/r 根据/p 自己/r 在/p 二战/j 转战/v 北非/ns ,/w 西欧/ns 期间/f 的/u 日记/n ,/w 在/p 战争/n 刚刚/d 结束/v 时/ng 撰写/v 的/u ,/w 也/d 是/v 他/r 本人/r 唯一/b 的/u 有关/vn 二战/j 的/u 连续性/n 记载/v 。/w 回顾/v 第二/m 次/q 世界大战/l ,/w 除了/p 战役/n 与/c 屠戮/v ,/w 还有/v 一些/m 重要/a 人物/n 。/w
整整/d 二十/m 世纪/n 。/w 美国/ns 一直/d 是/v 世界/n 第一/m 强/a 国/n 。/w 冷战/n 结束/v 后/f 美国/ns 所/u 占据/v 的/u 唯一/b 超级大国/n 地位/n ,/w 至少/d 可/v 以/p 保持/v 二三十/m 年/q 。/w 今天/t ,/w 美国/ns 是/v 中国/ns 最/d 重要/a 的/u 外交/n 对手/n ,/w 是/v 影响/v 中国/ns 经济/n 发展/vn 和/c 政治/n 桅顶/n 的/b 最/d 大/a 外部/f 力量/n ;/nx 同时/c ,/w 美国/ns 也/d 是/v 中国/ns 事实上/l 最/d 大/a 的/u 贸易/vn 伙伴/n ,/w 在/p 中国/ns 的/u 第一/m 大/a 投资国/n ,/w 是/v 同/p 中国/ns 在/p 教育/vn ,/w 科学/a 文化/n ,/w 技术/n 等/u 领域/n 交往/vn 最/d 多/a 的/u 国家/n ,/w 所以/c 我们/r 需要/v 借鉴/v 发达/a 资本主义/n 国家/n 的/u 经验/n ,/w 全面/ad 了解/v 美国/ns ,/w 深入/ad 研究/v 美国/ns 。/w

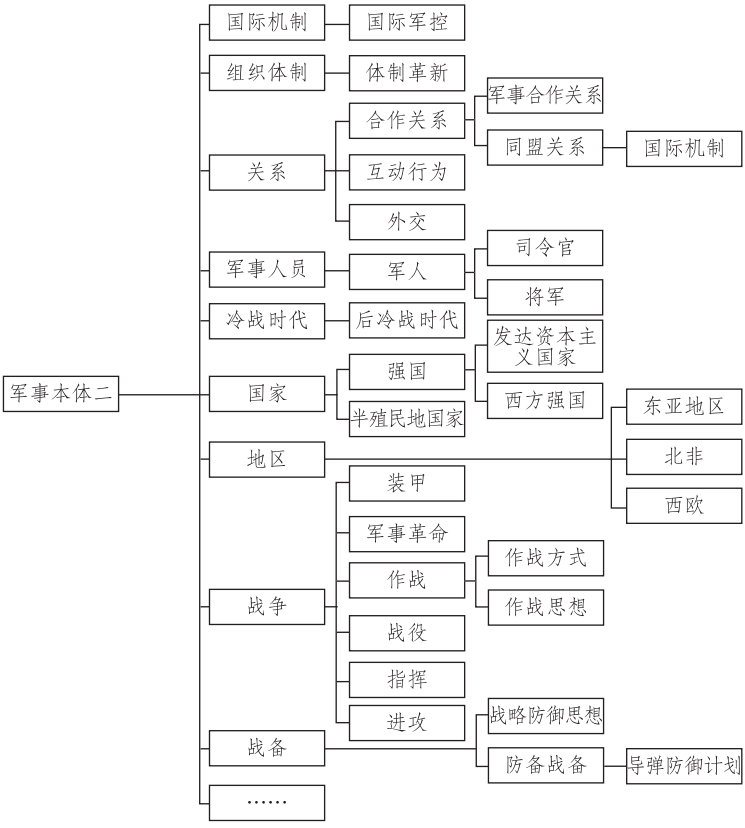

图 4　军事本体二类目等级体系(部分)

间的非等级关系分为类目之间的非等级关系、类目与实例间的非等级关系和实例之间的非等级关系。首先确定类目之间的非等级关系，利用关联规则的提取方法得到同级关系，比如“将军”和“司令”，利用语法规则的提取方法得到具有谓语语词的关系，比如＜半殖民地国家，跃升，强国＞，这样就可以确定类目的等级体系结构（见图4）。

再添加实例，在文本中，核心词汇之外的词均可能成为实例，为方便实验，本文将表3中非类目概念的词视为实例，因为实例无需从文本中再次提取，但要确定的是特定的实例应该隶属于哪个类，相当于要确定类目与实例之间的关系。可以利用相似度计算方法，将类目和实例间的两两相似度计算出来，相似度高的则认为某实例是某类目的实例，比如“中国”是“半殖民地国家”的实例，“第二次世界大战”是“战争”的实例，“巴顿”是“司令官”的实例等。需要说明的是，在有限的文本集合中，不是所有的实例都能找到对应的类目，也不是所有的类目都包含实例。

最后一种非等级关系是实例之间的关系，同样利用关联规则和语法规则的提取方法得到语词关系，比如＜美国，对手，中国＞，＜巴顿，擅长，追击＞等。

（5）形式化。除了像军事本体一利用本体语言进行形式化外，还可利用 Protégé 工具构建本体，如图5所示。

表3　基于 N-gram 且符合组词规则的概念提取结果（部分）

基于2-gram	组词规则	基于3-gram	组词规则
半殖民地国家	/b +/n	导弹防御计划	/n +/v +/n
导弹管制	/n +/v	国家安全战略	/n +/an +/n
竞争能力	/vn +/n	发达资本主义国家	/a +/n +/n
防务战备	/n +/vn	后冷战时代	/f +/n +/n
海岸警卫队	/s +/n	军事合作关系	/n +/n +/n
国际军控	/n +/n	战略防御思想	/n +/v +/n
合作关系	/n +/n	美俄关系	/j +/j +/n
作战方式	/v +/n	互动行为	/n +/ng +/v

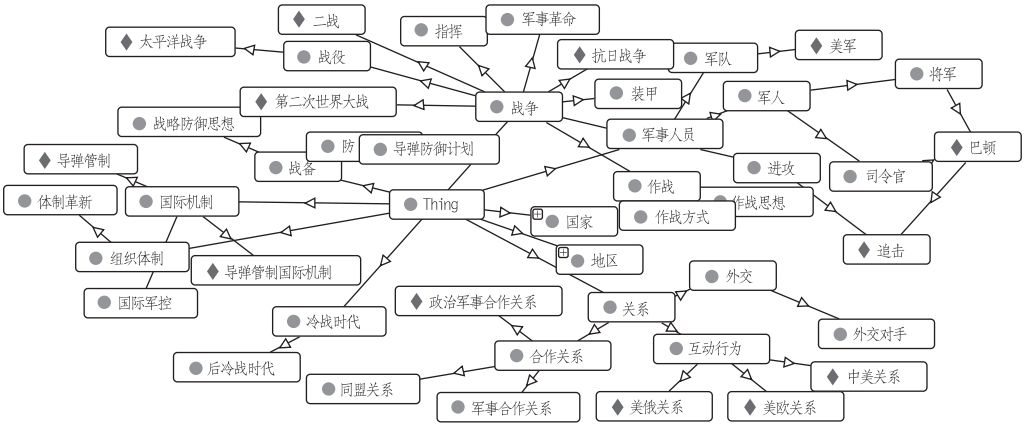

图5　军事本体二（部分）

4.3 讨论

从上述两个实验的过程和结果可以发现,馆藏资源语义化是需要一定条件的。首先,不同类型的图书馆要选择不同的语义化过程和关键技术。语义化过程是建立在已有馆藏资源数字化基础上的,目前有些小型高校图书馆、公共图书馆并不支持馆藏资源的公开检索和利用,因此离语义网还有一定距离,离馆藏资源的完全语义化则更远,这种图书馆首先应该进行完整的主题标引及规范的内外部特征信息组织。其次,馆藏资源的数字化丰富程度与关键技术的选择密切相关。实验中武汉大学图书馆的摘要收录较为完整,如果一个图书馆只数字化了图书的外部特征,语义化时则只需要按照实验一的流程进行,也就是说数字化丰富程度越高,选择的语义化关键技术就越多越复杂,这也是与馆藏资源建设阶段呈正比的。最后,关键技术的使用技巧是语义化人员培训时的重要内容。从实验二中看到关键技术并不是统统使用,而应针对不同数据特征进行筛选,总的来讲,不同数据源的外部特征语义化过程均与实验一类似,但关系提取时要注意没有引文的数据不能用共引分析;学科专业程度高的数据源在信息提取时选择基于词典和基于 N-gram 的方法较好较快,如实验二;特定表达形式的数据源则选择基于互信息和基于上下文信息的方法较准确,如诗歌、小说;而不论数据源专业程度如何,文本中的语义关系均很复杂,提取语义关系时,关联规则、聚类算法和语法规则需要结合使用。因此,这些关键技术需要经过语义化人员对数据和方法的深入理解后才能使用。

5 结语

馆藏资源语义化还需要一个长期的过程,本文从微观层面描述了馆藏资源语义化的全过程,设计了馆藏资源语义化模型,总结出馆藏资源语义化关键技术主要是信息提取相关技术、语义关系提取相关技术和形式化技术,分支技术则包括分词技术、命名实体识别、概念提取技术、相似度计算方法、聚类技术、计量分析技术、关联规则、形式化语言和工具等技术与方法。

本文利用武汉大学图书馆馆藏资源的不同特征分别进行了馆藏资源语义化模型的实验,对外部特征主要采用了关键技术中的相似度计算方法、计量分析技术、形式化语言和工具,对内部特征主要采用了分词技术、基于 N-gram 算法和组词规则结合的概念提取技术和基于语法规则和关联规则结合的关系提取技术等,分别验证了馆藏资源语义化模型中针对不同特征而设计的语义化流程。事实上,现实生活中的语义网不可能单独存在,不可能馆藏资源内部特征拥有一个语义网,外部特征拥有另外一个语义网,实验中的语义化内容其实是可以合并的,就是将内部特征语义资源整合入外部特征语义资源,或两者融合为整体的语义资源。因此馆藏资源语义化模型尚可改进,这是今后研究的方向。

参考文献

[1] 杨爱武. 基于关联数据的图书馆创新服务研究[J]. 图书与情报, 2012(3): 85-88. (Yang Aiwu. The research of library innovation service based on linked data[J]. Library and Informaion, 2012(3): 85-88.)

[2] 新浪科技. 路透社发布 Calais 网络服务开放式 API [EB/OL]. [2013-04-29]. http://tech.sina.com.cn/i/2008-01-31/14382008679.shtml. (Sina Technique. Reuters published a Calais Web services open API [EB/OL]. [2013-04-29]. http://tech.sina.com.cn/i/2008-01-31/14382008679.shtml.)

[3] 张海粟,马大明,邓智龙. 基于维基百科的语义知识库及其构建方法研究[J]. 计算机应用研究, 2011(8): 2807-2811. (Zhang Haili, Ma Daming, Deng Zhilong. Semantic knowledge bases construction based on Wikipedia[J]. Application Research of Computers, 2011(8): 2807-2811.)

[4] 夏翠娟,刘炜,赵亮,等. 关联数据发布技术及其实现[J]. 中国图书馆学报, 2012(2): 49-58. (Xia Cuijuan, Liu

Wei, Zhao Liang, et al. The current technologies and tools for linked data:A case of Drupal[J]. Journal of Library Science in China, 2012(2): 49 -58.)

[5] 乔晓东,白海燕,梁冰. NSTL 的关联数据构建与应用场景设想[J]. 数字图书馆论坛, 2012(2): 54 -60. (Qiao Xiaodong, Bai Haiyan, Liang Bing. Construction of linked data and design of application scenes in NSTL[J]. Digital Library Forum, 2012(2): 54 -60.)

[6] 戴维民. 语义网信息组织技术与方法[M]. 上海:学林出版社, 2008. (Dai Weimin. Information organization technology and method on semantic web[M]. Shanghai: Academia Press, 2008.)

[7] 李青山,陈平. 语义化互联网的关键技术[J]. 计算机科学, 2002(6): 86 -89. (Li Qingshan, Chenping. Research on key techniques of semantic web[J]. Computer Science, 2002(6): 86 -89.)

[8] 姚绍文. 语义化 web 的关键技术及其应用研究[D]. 成都:电子科技大学, 2002. (Yao Shaowen. Research on key issues and application of semantic web[D]. Chengdu: University of Electronic Science and Technology of China, 2002.)

[9] 代印唐. 基于语义网络的知识协作关键技术研究[D]. 上海:复旦大学, 2009. (Dai Yintang. Research on key technologies of semantic networks based on knowledge collaboration[D]. Shanghai: Fudan University, 2009.)

[10] 厉毅,郑炜. 数字学习网站资源的微格式语义化组织[J]. 中国教育信息化, 2012(17): 30 -33. (Li Yi, Zheng Wei. Digital learning website semantic organization based on micro formats[J]. China Education Info, 2012(17): 30 -33.)

[11] 罗庆云,赵巾帼. 语义化 Web 的理论基础与技术基础[J]. 甘肃联合大学学报(自然科学版), 2007(5): 75 -79. (Luo Qingyun, Zhao Jinguo. Semantics web rationale and technology base[J]. Journal of Gansu Lianhe University (Natural Sciences), 2007(5): 75 -79.)

[12] 朱大丽. 图书馆目录数据关联的语义化探析——充溢着背景知识的图书馆目录数据[J]. 图书馆学研究, 2012(1): 54 -58, 95. (Zhu Dali. Library catalog data linked semantization: Full of background knowledge library catalog data [J]. Research on Library Science, 2012(1): 54 -58, 95.)

[13] 白海燕,乔晓东. 基于本体和关联数据的书目组织语义化研究[J]. 现代图书情报技术, 2010(9): 18 -27. (Bai Haiyan, Qiao Xiaodong. Study of semantic bibliography base on ontology and linked data[J]. New Technology of Library and Information Science, 2010(9): 18 -27.)

[14] 欧石燕. 面向关联数据的语义数字图书馆资源描述与组织框架设计与实现[J]. 中国图书馆学报, 2012(6): 58 -71. (Ou Shiyan. Design and implementation of a linked data-oriented framework for resource description and organization in semantic digital libraries[J]. Journal of Library Science in China, 2012(6): 58 -71.)

[15] 王军,卜书庆. 网络环境下知识组织规范的研究与设计[J]. 中国图书馆学报, 2012(4): 39 -45. (Wang Jun, Bu Shuqing. A study and design on the standard for the networked knowledge organization system[J]. Journal of Library Science in China, 2012(4): 39 -45.)

[16] 王丽华. 基于语义网的数字图书馆的关键技术[J]. 情报杂志, 2004(4): 5 -8. (Wang Lihua. Key technology of digital library based on semantic web[J]. Journal of Information, 2004(4): 5 -8.)

[17] 王睿佳,刘耀. 面向科技文献的多模态语义关联特征提取与表达体系研究[J]. 大学图书馆学报, 2012(5): 71 -76. (Wang Ruijia, Liu Yao. Study on the feature extraction and expression system of multi-modal semantic information for scientific and technical literature[J]. Journal of Academic Library, 2012(5): 71 -76.)

[18] 董慧,余传明,姜赢,等. 基于本体的数字图书馆检索模型研究(Ⅱ)——语义信息的提取[J]. 情报学报, 2006(4): 451 -461. (Dong Hui, Yu Chuanming, Jiang Ying, et al. Research on the ontology-based retrieval model of digital library(II)—Semantic information acquisition [J]. Journal of the China Society for Scientific and Technical Information, 2006(4): 451 -461.)

[19] 刘炜. 基于本体的数字图书馆语义互操作[D]. 上海:复旦大学, 2006. (Liu Wei. Ontology-based semantic interoperability for digital libraries[D]. Shanghai: Fudan University, 2006.)

[20] 韩毅. 语义网格环境下数字图书馆知识组织策略与应用研究[D]. 长春:吉林大学, 2008. (Han Yi. Study on digital

library knowledge organization strategy and application under semantic grid environment[D]. Changchun: Jilin University, 2008.)

[21] 牟冬梅. 数字图书馆知识组织语义互联策略及其应用研究[D]. 长春:吉林大学, 2009. (Mou Dongmei. Study on semantic interconnection strategy and application on digital library knowledge organization[D]. Changchun: Jilin University, 2009.)

[22] 滕广青. 基于概念格的数字图书馆知识组织研究[D]. 长春:吉林大学, 2012. (Teng Guangqing. Research on knowledge organization based on concept lattice of digital library[D]. Changchun: Jilin University, 2012.)

[23] 董慧,余传明,徐国虎,等. 基于本体的数字图书馆检索模型研究(Ⅳ)——历史领域知识推理机制[J]. 情报学报, 2006(6): 666-678. (Dong Hui, Yu Chuanming, Xu Guohu, et al. Research on ontology-based retrieval model of digital library(Ⅳ)——Inference mechanism of history domain knowledge[J]. Journal of the China Society for Scientific and Technical Information, 2006(6): 666-678.)

[24] 刘成山,刘怀亮. 基于语义网的数字图书馆[J]. 情报杂志, 2008(1): 49-54. (Liu Chengshan, Liu Huailiang. Digital library based on semantic web[J]. Journal of Information, 2008(1): 49-54.)

[25] 贾保先,鲍素贞,杨吉宏. 虚拟数字图书馆语义平台建设关键技术研究[J]. 聊城大学学报(自然科学版), 2009(4): 93-96. (Jia Baoxian, Bao Suzhen, Yang Jihong. Research on key issues of virtual digital library semantic platform[J]. Journal of Liaocheng University (Natural Science Edition), 2009(4): 93-96.)

[26] 董慧,杨宁,余传明,等. 基于本体的数字图书馆检索模型研究(Ⅰ)——体系结构解析[J]. 情报学报, 2006(3): 269-275. (Dong Hui, Yang Ning, Yu Chuanming, et al. Research on the ontology-based retrieval model of digital library (Ⅰ)—Explanation of the architecture[J]. Journal of the China Society for Scientific and Technical Information, 2006(3): 269-275.)

[27] 董慧,余传明,杨宁,等. 基于本体的数字图书馆检索模型研究(Ⅲ)——历史领域资源本体构建[J]. 情报学报, 2006(5): 564-574. (Dong Hui, Yu Chuanming, Yang Ning, et al. Research on the ontology-based retrieval model of digital library(Ⅲ)——History domain ontology building[J]. Journal of the China Society for Scientific and Technical Information, 2006(5): 564-574.)

[28] 潘伟. 个性化信息服务的关键技术——聚类分析[J]. 现代情报, 2007(10): 212-214. (Pan Wei. Personalized information service key technologies—cluster analysis[J]. Modern Information, 2007(10): 212-214.)

[29] 李静. 数据挖掘技术在高校图书馆个性化服务中的应用研究[D]. 天津:天津大学, 2012. (Li Jing. Study and application of data mining technology in personalized service of the university libraries [D]. Tianjin: Tianjin University, 2012.)

[30] 赵红霞. 数据挖掘技术和RSS技术在图书馆个性化服务中的应用[D]. 郑州:解放军信息工程大学, 2008. (Zhao Hongxia. Data mining technology and RSS technical on electronic library application[D]. Zhengzhou: The PLA Information Engineering University, 2008.)

[31] 周庆. 图书馆个性化信息服务的技术支持[J]. 大学图书情报学刊, 2008(6): 60-64. (Zhou Qing. The technology support to the individualized information service in libraries[J]. Journal of Academic Library and Information Science, 2008(6): 60-64.)

[32] 张炜,洪霞. 基于OPAC读者行为挖掘的个性化服务系统关键技术分析[J]. 图书馆论坛, 2010(1): 62-64. (Zhang Wei, Hong Xia. The analysis of key technology in individual service system based on the OPAC reader behavior excavation[J]. Library Tribune, 2010(1): 62-64.)

[33] 王思丽,祝忠明. 利用关联数据实现机构知识库的语义扩展研究[J]. 现代图书情报技术, 2011(11): 17-23. (Wang Sili, Zhu Zhongming. Study on the semantic expansion of institutional repository based on linked data[J]. New Technology of Library and Information Science, 2011(11): 17-23.)

[34] 贺德方,曾建勋. 基于语义的馆藏资源深度聚合研究[J]. 中国图书馆学报, 2012(4): 79-87. (He Defang, Zeng

Jianxun. Study on in-depth integration of library collections based on semantics[J]. Journal of Library Science in China, 2012(4): 79 -87.)

[35] 邱均平,余凡. 基于计量分析的馆藏资源语义化理论研究[J]. 中国图书馆学报, 2012(4): 71 -78. (Qiu Junping, Yu Fan. Theoretical research on semantization of library resources based on informetric analysis [J]. Journal of Library Science in China, 2012(4): 71 -78.)

[36] 邱均平,楼雯. 基于共现分析的语义信息检索研究[J]. 中国图书馆学报, 2012(6): 89 -99. (Qiu Junping, Lou Wen. Semantic information retrieval research based on co-occurrence analysis [J]. Journal of Library Science in China, 2012(6): 89 -99.)

[37] 符福垣,吴显沪. 情报科学的基本概念(三)[J]. 情报科学, 1985(6): 72. (Fu Fuyuan, Wu Xianhu. Basic concepts of information science(Ⅲ)[J]. Information Science, 1985(6): 72.)

[38] 袁璐,蒙祖强,许珂. 依存分析和 HMM 相结合的信息抽取方法[J]. 计算机工程与应用, 2012(9): 138 -141. (Yuan Lu, Meng Zuqiang, Xu Ke. Method of text information extraction based on dependency parsing and HMM[J]. Computer Engineering and Applications, 2012(9): 138 -141.)

[39] 熊回香,夏立新. 汉语分词技术综述[J]. 图书情报工作, 2008(4): 81 -84. (Xiong Huixiang, Xia Lixin. The review of Chinese automatic word segmentation technology [J]. Library and Information Service, 2008(4): 81 -84.)

[40] ICTCLAS 汉语分词系统[EB/OL]. [2013 -05 -01]. http://ictclas.org/index.html. (ICTCLAS Chinese word segmentation system[EB/OL]. [2013 -05 -01]. http://ictclas.org/index.html.)

[41] 分词算法[EB/OL]. [2013 -05 -02]. http://blog.csdn.net/cozmic/article/details/659591. (Segmentation[EB/OL]. [2013 -5 -2]. http://blog.csdn.net/cozmic/article/details/659591.)

[42] 丁卓冶. 中文命名实体识别的研究[D]. 大连:大连理工大学, 2008. (Ding Zhuozhi. A study on Chinese named entity recognition[D]. Dalian: Dalian University of Technology, 2008.)

[43] 潘家铭. 基于 Wikipedia 的中文命名实体识别研究[D]. 广州:中山大学, 2008. (Pan Jiaming. A study on Chinese named entity recognition based on Wikipedia [D]. Guangzhou: Sun Yat-sen University, 2008.)

[44] Velardi P, Missikoff M, Basili R. Identification of relevant terms to support the construction of domain ontology[C]// Proceedings of the workshop on Human Language Technologies and Knowledge Management. Stroudsburg, PA: Association for Computational Linguistics, 2001: 1 -8.

[45] Tversky A. A feature of similarity[J]. Psychological Review, 1977, 84(4): 327 -352.

[46] Valerie C. Fuzzy semantic distance measures between ontological concepts [J]. IEEE Annual Meeting of the Fuzzy Information, 2004: 635 -640.

[47] Rodriguez M A, Egenhofer M J. Determining semantic similarity among entity classes from different ontologies[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(2): 442 -456.

[48] Macqueen J. Some methods for classification and analysis of multivariate observations[C]// Lucien M, Le C, Jerzy N. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967, 281 -297.

[49] Ng R, Han J. Efficient and effective cluster method for spatial data mining[C]// Bocca J, Darke M, Zanio C. Proceedings of the 20th International Conference of Very Large Data Bases. San Francisco, CA: Morgan Kaufmann Publisher, 1994: 144 -155.

[50] Guha S, Rastogi R,Shim K. CURE: An efficient clustering algorithm for large databases[C]// Laura M H, Ashutosh T. Proceedings of the ACM SIGMOD Conference, Seattle, Washington: ACM Press, 1998: 73 -84.

[51] Ester M, Kriegel H P, Sander J, et al. Adensity-based algorithm for discovering clusters in large spatial databases with noise[C]// Evangelos S, Jiawei H, Usama M F. Proceedings of the 2nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Portland, Oregon: AAAI Press, 1996: 226 -231.

英法两国分别实施网络资源法定呈缴制度

英国自2013年4月6日起正式实施电子出版物法定呈缴制度。电子书、电子期刊以及可被存储在 CD-ROM 和从网站下载的电子出版物等网络电子资源,将被送存至大英图书馆为首的六家图书馆,以实施对国家文化和数字格式内容的收集和保存。

法国在2006年8月1日通过的法国文化遗产规章(Code du patrimoine)规定,法定呈缴制度的覆盖范围延伸至互联网领域。法国国家图书馆可以对法国境内的网站资源进行采集、保存,并向公众开放,出版商不能阻挠图书馆的采集工作,并应该向法国国家图书馆提供在线资源。根据2012年12月公布的新修法令,法国国家视听研究院(Inathèque de France)负责采集与视听通信相关的网站(以广播电视为主),法国国家图书馆负责采集其他所有类型的网站,收集"任何在法国出版(或进口)"的资源。2013年8月法国国家图书馆已发布工作进展以及问题释疑。

显然,英法两国的实施路径有别:英国由六家图书馆联合实施将呈缴范围扩大到"电子出版物"的法定呈缴制度,法国则是将网络典藏归属在文化遗产保护的范畴中,对包括互联网网站内容在内的"数字资源"进行采集。英法两国国家图书馆的网络资源典藏制度,对欧盟乃至世界范围的国家图书馆推动相关工作具有引领作用。

资料来源

1. Introduction to legal deposit. http://www.bl.uk/aboutus/legaldeposit/introduction/index.html.
2. Legal deposit for websites and electronic publications. http://www.bl.uk/aboutus/legaldeposit/websites/index.html.
3. Click to save the nation's digital memory. http://pressandpolicy.bl.uk/Press-Releases/Click-to-save-the-nation-s-digital-memory-61b.aspx.
4. Qu'est-ce que le dépôt légal? http://www.bnf.fr/fr/professionnels/depot_legal_definition/s.depot_legal_mission.html.
5. Digital legal deposit. http://www.bnf.fr/en/professionals/digital_legal_deposit.html.

(顾立平 姚伟欣 张舵 整理)

[52] Wang W, Yang J, Muntz R. STING: A statistical information grid approach to spatial data mining[C]// Matthias J, Michael J C, Klaus R D, et al. Proceedings of the 23rd Conference on VLDB. Athens, Greece: Morgan Kaufmann, 1997: 186-195.

[53] Chang Rui, Liu Zhiyi. An improved Apriori algorithm[C]//Proceedings of 2011 International Conference on Electronics and Optoelectronics. Washington, DC: IEEE Computer Society, 2011: 476-478.

楼 雯 武汉大学信息管理学院博士研究生。通讯地址:武汉大学信息管理学院。邮编:430072。

(收稿日期:2013-05-03)