ICS 01.140.20
A 14

C A D A L 项 目 标 准

CADAL 41101—2012

数字图书馆知识元抽取规范

Digital Library Knowledge Indexing Standards—
Digital Library Knowledge Element Extraction Standard

第一稿

2012-05-08 发布　　2012-05-09 实施

CADAL 项目管理中心　　发　布

目　　次

前　言

《数字图书馆知识标引标准规范》分为 3 个部分：

——第 1 部分：数字图书馆知识元抽取规范；

——第 2 部分：数字图书馆资源学科分类标准；

——第 3 部分：数字图书馆学科文献学术水平等级切分标准。

本规范为其中的第 1 部分。

CADAL 数字图书馆知识元抽取规范是依据中美双方对 CADAL 内容的共同要求，参照国家科技部科技基础研究重大科技专项“我国数字图书馆标准与规范建设”项目之《数字图书馆集成服务描述标准发展现状》、《知识组织体系应用标准规范研究》等研究成果，对数字图书馆的知识元抽取技术进行分析而制定的。

本规范由大学数字图书馆国际合作计划(CADAL)项目管理中心归口。

本规范起草单位：宁波大学图书馆。

本规范主要起草人：刘柏嵩、董其军、毛海波、豆洪青。

引　言

知识组织是在传统文献信息环境下发展起来的信息组织和利用手段，在百余年的应用过程中，形成并完善了分类法、主题词表等知识组织工具。在今天的网络环境下，知识组织工具需要进一步发展和创新，以适应网络化的信息获取手段，满足数字化信息资源组织的要求。随着网络信息技术的发展，数字图书馆的内涵也正在发生着根本性的改变，它不再是单个的系统，而变成了一个开放的数字环境，即一种基于知识内容、应用环境和应用群体有机交互的数字化知识化服务机制。在这种泛在的数字环境下，如何将分散的、异构的数字资源整合在一起，成为有组织的整体，使之能够有效地被保存、发现和获取，从而构建一种新的服务模式，成为新一代数字图书馆研究的重要课题。国内外学者提出了构建一种新型数字图书馆运行模式，通过建立通用的信息资源语义描述机制，实现分布异构的数字资源的整合，从而实现面向服务的数字图书馆运行模式：面向机器的服务和面向用户的服务。面对互联网这个开放的海量资源环境时，借助机器的方式，是网络信息进行有效组织和控制的必需路径。利用机器来进行信息发现和知识组织的基础是机器理解。只有机器能够认识信息的特征，理解信息的含义才有可能对信息进行组织和发现。本规范以机器能自动或半自动形式进行数字图书馆的资源组织与标引为目标，提供一种通用的、形式化的知识标引描述框架。

图书馆服务在数字时代面临的一个巨大的挑战是如何深入到更细小的知识单元(如研究方法、原理和数据)，进行组织、整理、策管(curator)和服务，而不局限于电子书、期刊文章、技术报告等。由此可实现新的技术架构(包括关联数据、知识组织、云平台和移动技术等)，让虚拟图书馆逐渐走向后台，隐形于各类网络服务之中，不一定要直接面向读者，而是作为一种基础服务(包括数据服务)，成为赛百空间的基础设施之一。这种新的存在形式，真正能够体现数字图书馆的价值，特别是能够对科研、教育和医药卫生等方面提供持续的支持。

知识元标引可使知识被有效地检索、利用，实现知识创新和增值，为用户提供针对性的知识服务，能很好地解决以上问题。知识元是知识的最小单位，以知识元为单位的知识标引为用户提供的不再是文献，而是文献中的具体知识，在一定程度上满足了人类对知识组织、知识管理、知识服务的需求。

随着信息服务转向知识服务，信息资源管理向知识资源管理转变，信息组织向知识组织转变，知识的组织结构由等级式向网络式转变，以文献为单位的传统标引也要向以知识为单位的知识标引转变。现有的知识组织方式组织的是知识的载体——文献，而非知识本身；只能保证检出的文献含有所需知识，而不能揭示这些知识之间的联系，不能为产生新知识提供联系。可通过标题、关键词、作者、内容分类特征等元数据进行关联检索，然而对于用户所要解决的问题来说，却不能提供全面、快速、准确的知识信息。通过本规范的制定，对知识元标引的各环节进行规范，指导数字图书馆知识标引工作发展。

现有的分类或主题词表不强调组织概念关联，没有实现语义层面的资源标注，因而无

法准确、完整地显示资源的知识结构，所组织的知识也不能以知识网络方式显示。本规范通过关联数据加强资源体系中概念之间的关联，并强调基于语义的资源标注。本规范提出采用基于语义的关联开放数据方式，对数字资源进行知识标引。基于关联数据的知识序化与控制，即是在分析知识所属领域的实体对象关联关系中提取大量的新知识，并对其进行分析与综合，形成新的知识关联，从而生产出更高层次上的综合的知识产品。其目的是改变知识因子间的原有联系，其结果可以提供新知识。

本规范是根据 CADAL 项目“知识组织框架”规范要求，参考现行的知识标引和知识组织著录规范而制定的。本规范的适用对象包括中文图书、西文图书、期刊及学位论文等印刷型文献的数字化文本以及原生数字资源。

数字图书馆知识元抽取规范

1 范围

本规范规定了数字图书馆知识标引(CADAL Knowledge Indexing，CKI)中知识元抽取的总体要求和方法。

本规范适用于文献资源数据库服务平台。

2 规范性引用文件

数字图书馆标准规范发展战略子项目组. 我国数字图书馆标准规范建设与应用的实施指南. [2012-04-05]. http: //cdls2. nstl. gov. cn/2003/Whole/TecReports. html.

潘淑春，盛玲玉，牛离平. 数字图书馆相关领域标准规范现状与发展研究(数字科研). [2012-04-05]. http: //cdls2. nstl. gov. cn/2003/Whole/TecReports. html.

富平，等. 数字图书馆知识组织体系标准规范应用机制研究 . [2012-04-05]. http: //cdls2. nstl. gov. cn/2003/Whole/TecReports. html.

韩松涛，等. CADAL 多维度标签分类标准. [2011-09-28].

3 术语与定义

3.1 CKI 知识元

知识元，是指不可再分割的具有完备知识表达的知识单位，包括概念知识元、事实知识元和数值型知识元等。具体指文献资源中作为知识输入、知识创新及知识输出的单元，包括理论、原理、概念、定义、范例、规则和结论等，表现为一种基于语义的关联开放数据(semantic based linked open data)。

3.2 关联数据

关联数据是一种数据发布方式。根据关联数据原则，关联数据技术的应用是为了方便实现数据集、元素集及词汇集之间的关联。关联数据使用统一资源标识符(uniform resource identifier，URI)作为唯一标识符标识任何类型的资源，类似于传统图书馆领域如何使用规范控制的标识符。关联数据技术使用统一的标准描述数据(resource description framework，RDF)，可以明确各实体之间的关系；各实体之间的关系可以用于导航、资源整合。

3.3 知识本体

知识本体是领域知识的形式化说明，通常由概念、概念之间的关系、公理和规则组成。

3.4 CKI 知识元标引

将知识视为独立于文献的元素进行标识，是实现跨领域知识集成与知识发现的基础。深入挖掘知识元之间隐藏的关系，进而发现新知识、创造新领域。

3.5 标引方式

自动标引即计算机辅助标引，是根据文献内容，依靠计算机系统全部或部分地自动给出标引符号的过程。

3.6 知识组织

对知识元进行标引并构建其语义关系。

3.7 知识关联

知识关联主要是通过知识元链接和引文链接将文献间的知识关联起来。通过文献之间、知识元之间、分类导航之间的交叉链接，构建起节点丰富、交织纵横的知识网络系统。知识元链接包括作者、机构、刊名、关键词、相关作者群、相关研究机构、相关关键词等。

4 CKI 知识元抽取导论

4.1 CKI 知识元的内容对象基本准则

(1)知识元是显性知识(explicit knowledge)的最小可控单位。

(2)知识元是完备的，即一个知识元在逻辑上是完整的，能表达一个完整的事实、原理、方法、技巧等。

(3)使用 URI 来标识知识元；使用 HTTP URI 使人们可以访问到这些标识；当有人访问到标识时，提供有用的信息。

(4)尽可能提供关联的 URI，以使人们可以发现更多的事物；每个数据库会公布所采用的 RDF 词表，以便消费程序理解数据的格式。

(5)众多的知识元通过一定的语义连接在一起，可以实现知识价值的增值，甚至催生新的知识。除数据本身的属性之外，不同资源之间的语义关联能被揭示出来。

4.2 CKI 知识元的语义模型

知识元由以下三个层次构成：检索层次，包括学科专业名称、分类号、知识元名称、知识级别、知识元关键词等检索点；知识元描述层次，对知识元的内容进行完整的文字描述及声像辅助描述；关联层次，与该知识元相关联的内容链接条目清单，由知识元链接地址和知识元关键词之间的语义关系进行扩展。

由此，CKI 知识元可定义为一个六元组：CKIE=(KEID，N，C，KC，KL，L)。

式中：

KEID——CKI 知识元标识号，采用 URI 表示。

N——知识元名称，即知识元的标题，是对本知识元的知识内容的一种高度概括表述，用一个较短的字符串表示。

C——知识元的关键字集，即一组概念集，可用于检索本知识元的关键字集。

D——简要说明，是对本知识元的内容的简要描述。通过该描述，一方面可以使在构建知识元结构模型时选择知识元的准确性提高；另一方面，在实现知识共享时，可以使知识使用者直接获得知识内容，而不需要再到庞大的载体中去寻找。

KC——知识类别，指对知识体系按照学科标准分成若干领域，并且与学科分类号分别对应；对知识体系按照一定标准划分后的各个子项，如采用学科为标准，知识类别可以分为生物、哲学、管理学、文学等领域，表现为层次关系。

KL——知识级别，按照认知方式分成概念、公理、规则和方法四类；知识级别由简单到复杂包括概念、公理、规则和方法四个层次。概念指在头脑里所形成的反映对象的本质属性的思维形式；公理指依据人类理性和愿望发展起来而共同遵从的道理；规则指长期形成的规律；方法指为达到某种目的而采取的途径、步骤、手段等。从概念、公理、规则到方法反映了人们由浅到深的认知方式。文献资源中知识元的知识级别，也可作为该文献学术水平切分的一个参考依据。

L——知识地址，指在构建知识库时所赋予某一知识元的唯一位置标识，一般是个链接到知识元所在的载体的超级链接。通过该信息，知识的使用者可以获得对该知识元更深入完整的信息。

以知识元为结点，以知识元之间的关联(包括知识元的地址关联、知识元关键词的语义关联等)为边建立起来的链接称为知识元网。该网络可以表示为 KENet=(KE，R)，其中，KE=$\{ke_1, ke_2, \cdots, ke_n\}$为知识元集合，R 为知识元之间关联关系的集合。其中，关键词(概念)之间的语义关联定义如下：

定义(层级关系)：根据概念间的包含关系，可将概念区分为上位概念和下位概念。上位概念称为大概念，下位概念称为小概念。按同一标准(同一维度)划分并处于同一层面的概念称为并列概念。它主要指属种关系，即概念外延的包含关系。小概念(种)的外延是大概念(属)外延的一部分。小概念除了具有大概念的一切特征外，还具有本身独有的区别特征。示例：(属)—树；(种)—乔木、灌木。

定义(父结点)：在知识元 O，$H=(h, <)$中，如果 x，$y\in h$，$x<y$，且不存在概念($z\neq x$，y)，满足 $x<z$ 且 $z<y$，则概念 y 称为概念 x 的父结点。

定义(规则分类体系)：若知识元概念分类体系 $H=(h, <)$满足以下两点，则称 H 是规则的：

(1)h 中存在一最大概念 y，对任意的 $x\in h$，有 $x<y$；

(2)存在概念集 $h_i(i=0, 1, \cdots, n-1)$，使得 $h=\bigcup_{i=0}^{n-1} h_i$ 且 $h_i \cap h_j=\varnothing (i\neq j)$；

(3)若 h_i 中一个概念的父节点在 h_j 中，则 h_i 中所有概念的父节点都在 h_j 中($i\neq j$)。

定义(超类、子类)：对任意的类 C_1 和 C_2，如果 $\forall i$：是实例(i，C_2)→是实例(i，C_1)，

即 $domain(C_2) \subseteq domain(C_1)$，则称 C_2 为 C_1 的子类，记为是子类(C_2，C_1)；相应地，C_1 称为 C_2 的超类，记为是超类(C_1，C_2)。

为便于讨论，我们引入以下关系：

(1)有子类(C；C_1，…，C_n)≡是子类(C_1；C)∧… ∧是子类(C_n；C)；

(2)是子类(C；C_1，…，C_n)≡是子类(C；C_1)∧… ∧是子类(C；C_n)；

(3)有超类(C；C_1，…，C_n)≡是子类(C；C_1)∧… ∧是子类(C；C_n)≡是子类(C；C_1，…，C_n)；

(4)是超类(C；C_1，…，C_n)≡是子类(C_1；C)∧… ∧是子类(C_n；C)≡有子类(C；C_1，…，C_n)。

在一个类的子类中可能存在某种次序，这种次序可能来源于领域的约定，也可能是便于记忆和教学。为此，引入有顺序子类(C；C_1，…，C_n)。它逻辑上与有子类(C；C_1，…，C_n)等价，但是保留了领域的子类出现次序：

有顺序子类(C；C_1，…，C_n)≡有子类(C；C_1，…，C_n)。

定义(直接超类、直接子类)：对任意的类 C_1 和 C_2，C_1 不同于 C_2，并且是子类(C_2，C_1)。如果对任意的类 C_3，是子类(C_3，C_1)∧是子类(C_2，C_3)→$C_2=C_3$，则称 C_1 是 C_2 的直接超类，记为是直接超类(C_2，C_1)；称 C_2 是 C_1 的直接子类，记为是直接子类(C_1，C_2)。

在分类体系中，一个类可以是多个类的直接子类，或者说一个类可以有多个直接超类。

定义(类的覆盖与划分类)：C，C_1，…，C_n 为知识元概念中的类，并且 C_1，…，C_n 为 C 的子类，如果∀i：是实例(i，C)↔∃k：是实例(i，C_k)，则称{C_1，…，C_n}是 C 的一个覆盖。如果{C_1，…，C_n}中的任何真子集不是 C 的覆盖，则称{C_1，…，C_n}是 C 的极小覆盖。如果{C_1，…，C_n}是 C 的极小覆盖，并且∀i，∀j：$i \neq j \rightarrow domain(C_i) \cap domain(C_j) = \varnothing$，则称{$C_1$，…，$C_n$}为 C 的一个分类(classification)或划分(partition)。

根据超类和子类关系，可以定义类的等价。

定义(类的等价)：对任意的类 C_1 和 C_2，类 C_1 等价于 C_2，记为 $eqv(C_1, C_2)$ 当且仅当是子类(C_2；C_1)∧是子类(C_1；C_2)，或者是超类(C_1，C_2)∧是超类(C_2，C_1)。

定义(上位节点)：给定一分类系统，其中 C，C_1，…，C_n 为其节点。如果∀i：是成员(i；C_i)→是成员(i，C)，则 C 称为 C_1，…，C_n 的上位节点，记为是上位节点(C；C_1，…，C_n)。

分类关系主要考虑上下位关系(Is-A)和部分整体关系(Part-Of)。

定义(Is-A)：对于概念集 SC 中的概念 C_1，$C_2 \in S_C$，如果有：①概念 C_1 的内涵包含 C_2 的内涵，即 $I(C_1) \supset I(C_2)$；②概念 C_1 的外延包含于 C_2 的外延，即 $E(C_1) \subset E(C_2)$。该关系存在于种概念和类概念之间。$A\ Is\text{-}A\ B =_{def} \forall x(inst(x, A) \rightarrow inst(x, B))$。则将概念 C_1 和 C_2 之间的关系称为种属关系，记作 $Is\text{-}A(C_1, C_2)$。

定义(Part_Of)：知识元概念 A 与 B 是部分关系，当且仅当：对 A 的任一实例 x，存在 B 的某些实例 y 在实例级与 x 为部分关系；反之亦然，即定义为 $A\ Part\_Of\ B =_{def} A part\_forB \& Bhas\_partA$。$Part\_Of$ 关系不具自反性和对称性，但具传递性。

在 CKI 知识元(概念)中，除了上述的分类关系外，还包括非分类关系(non-taxonomic)，又称为非层级关系，它反映了概念间的某些语义关系，其类型多种多样，主要

包括：序列关系、空间关系、时间关系、因果关系、源流关系、发展关系、联想关系、推理关系以及整体部分关系等。

5 知识元抽取

知识元是对一个知识的完整描述。可以是一个独立的学科知识单元，也可以是一个事物的过程或结果、结论。知识元抽取按照自顶而下的方式，将知识按学科层次划分成多个层次的模块及其相应的知识子模块，每个知识子模块构成一个知识单元，即为知识元。

5.1 知识元的内容构成

知识元包括以下几个方面的内容：

(1)所属学科类别、关键词、分类号。有两个作用：一是说明该知识元在学科模块层次中的位置，可作为类别检索点和“点击”快速返回某一学科层次；二是用于实现与关联内容自动链接的链接点。

(2)知识元名称。名称是知识元的唯一直接标识，以相应的标准主题词命名，是直接检索该知识元的一个检索途径。

(3)知识元描述。对知识元的内容进行完整的文字描述及声像辅助描述。该描述应包括公理、公式、定义、推论、事实、事件、事例、数表等内容，并随着科研的进步不断更新。

(4)关联条目。与该知识元相关联的内容链接条目清单。其内容包括：相关参考书刊，与其他学科、知识元的交叉内容，相关多媒体声像内容，相关网络信息内容等关联信息。网络信息包括：学术期刊、专利等专题数据库，相关专业技术、企业产品信息网站，市场社会需求，专家网上解答等。

关联条目的格式内容包括关联条目所属学科类别、题名或知识元名称、关键词、分类号，用于与知识元实现自动链接的链接点。

5.2 知识元抽取的方法

知识元的抽取可根据类型不同，采取不同的方法。知识可分为以下几种类型：①概念类知识元，是对事物性质、事物变化规律的认识，如“杠杆平衡”是一个概念。②原理类知识元，是对事物性质、事物变化规律的认识，如“杠杆平衡原理”是一个原理。③方法类知识元，解决同样的问题，方法可以多样，方法类知识元是指分析、解决问题的某种确定的方法，如“因式分解法”有配方法、十字相乘法、求根法等。④事实类知识元，反映一个事实，如历史事件、地理现象、社会现象等。⑤陈述类知识元，是用来表述两者之间的关系或为了表达某个观点，如生物学的基本特征、细胞中的种类和含量等。⑥数值类知识元，是用来表述对象或过程的数量特征和关系，如工业总产值、变化量、变化率等。⑦模型类知识元，用来描述事物或对象的数学或图形模型，如统计模型、双螺旋结构等。知识元抽取框架如图 1 所示，CKI 知识元抽取系统实现流程如图 2 所示。

示例 1：事实型知识元抽取系统框架 CKI (D, T)

/* 在数字资源集文档 D 的基础上抽取知识元 T */

Step1: Docs=preprocess(POS-tag(D))// 语料库(corpus)和 Web 文档集的收集、选择和预处理；

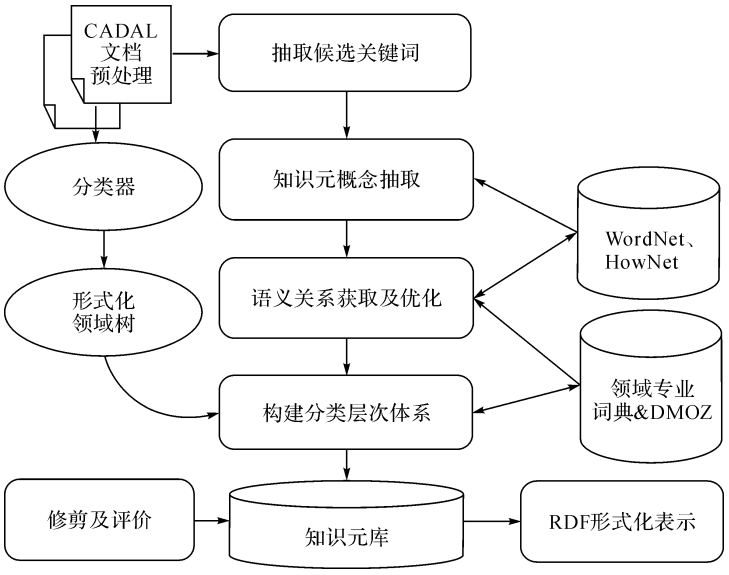

图 1　知识元抽取框架

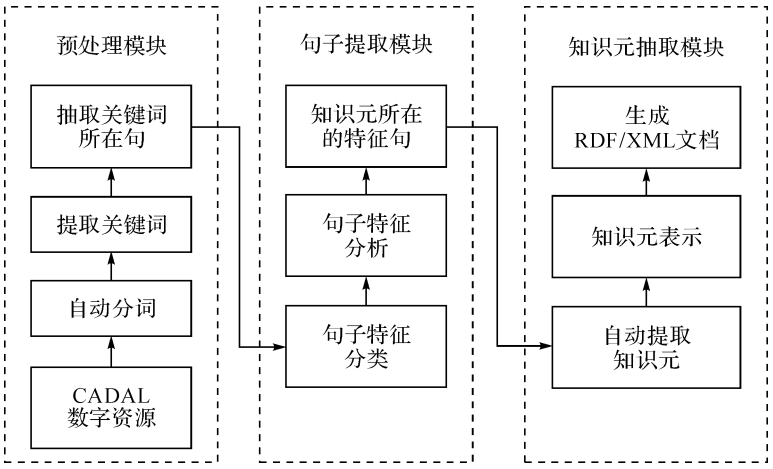

图 2　CKI 知识元抽取系统实现流程

Step2：Parses＝parse(POS-tag(D))// 生成候选关键词集；

Step3：DomainTerm＝Linguistic(Parses)&Statistical(Parses)// 抽取领域术语，采用语言学分析和统计分析的方法；

Step4：OntConcept＝Disambiguate(DomainTerm)// 对领域本体概念(concept)进行语义排歧；

Step5：R_{can}＝{Tem_{hyp}，Tem_{par}，Tem_{syn}，Tem_{ins}}// 候选语义关系(semantic relation)学习；

Step6：T_c＝AssociateRule(R_{can})// 语义关系优化；

Step7：DomainTree＝getFormalContext(T_c) // 创建形式化领域树(Domain Trees)；

Step8：T_o＝TaxMiner(DomainTree) // 构建分类层次体系(taxonomy)；

Step9：O＝Prune(T_o)//生成知识元和形式化表示；

Step10：Evaluate(KEU)//对生成的知识元进行评价；

Step11: Return KEU

示例 2：CKI 知识元概念的语义关系抽取——基于聚类的分类关系学习

输入：知识元概念列表 T

结果：计算每一概念对之间的相似度，并将其从由高到低进行排序

Step1. 初始化概念聚类集 C，i. e. C：={}

Step2. 对每一对 pair(t_1, t_2)，如果 t_1 或 t_2 没有分类为某一其它概念的子概念：

(a) IF (t_1, m)$\in H(t_2)$

i. (t_2, n)$\in H(t_1)$ and $n>m$, then isa(t_1, t_2)

ii. ELSE $isa(t_2, t_1)$

(b) ELSE IF (t_2, m)$\in H(t_1)$

i. $isa(t_2, t_1)$

(c) ELSE IF (h, n)$\in H(t_1)$ and (h, m)$\in H(t_2)$ and there is no h' such that (h', p)$\in H(t_1)$and (h', q)$\in H(t_2)$ and $p+q>m+n$

i. IF $isa(t_1, t')$, i. e. t_1 已分类为 t'

A. IF $t'==h$, then $isa(t_2, t')$

B. ELSE IF (h, n)$\in H(t')$ and ((t', m)$\in H(h)\rightarrow m<n$)

IF t_2 has not yet been classified, then $isa(t_2, t')$

IF t' has not yet been classified, then $isa(t', h)$

C. ELSE

IF t_2 has not yet been classified, then $isa(t_2, h)$

IF h has not yet been classified, then $isa(h, t')$

ii. ELSE IF $isa(t_2, t')$, i. e. t_2 is already classified as t'

A. IF $t'==h$, then $isa(t_1, t')$

B. ELSE IF (h, n)$\in H(t')$ and ((t', m)$\in H(h)\rightarrow m<n$)

as t_1 has not yet been classified, then $isa(t_1, t')$

IF t' has not yet been classified, then $isa(t', h)$

C. ELSE

as t_1 has not yet been classified, then $isa(t_1, h)$

IF h has not yet been classified, then $isa(h, t')$

iii. ELSE, as neither t_1 nor t_2 have been classified, $isa(t_1, h)$, $isa(t_2, h)$

(d) ELSE, as there are no common hypernyms, mark t_1 and t_2 as clustered, i. e. $C:=C\bigcup(t_1, t_2)$

Step3. 对每一概念 $t\in T$，由于在语料中未找到相似概念而未处理，如果 C 中有其他概念 t'满足 $stringOf(t', t)$, then $isa(t, t')$

Step4. FOR EACH(t_1, t_2)$\in C$

(a) IF there is a t' such that $isa(t_1, t')$ THEN $isa(t_2, t')$

(b) ELSE IF there is a t' such that $isa(t_2, t')$ then $isa(t_1, t')$

(c) ELSE select the pair(t', m)$\in H(t_1)\bigcup H(t_2)$ for which there is no (t'', n)$\in H(t_1)\bigcup H(t_2)$ such that $n>m$ and create the following structures: $isa(t_1, t')$ and $isa(t_2, t')$

Step5. FOREACH term $t\in T$ which has not been classified, put it directly under the top concept, i. e. $isa(t, top)$

Step6. 输出：知识元概念列表 T 的概念层次

CKI 数据模型将知识组织系统视为由概念集合组成的概念体系(concept scheme)。CKI 概念体系和 CKI 概念用 URIs 来辨识，使得任何人在任何上下文环境中可以一致地引用它们，将它们作为万维网的一部分。CKI 概念可以使用任意数量的词汇字符(例如“romantic love”或“れんあい”)，任意指定自然语言(例如中文、英语或日语平假名拼法)作为其标签。指定语言的标签中的一个可以作为该语种的首选标签，其他作为可选标签。CKOS 概念通过语义关系属性与其他 CKOS 概念关联起来。CKOS 数据模型提供 CKOS 概念间的等级和相关链接。

CKI 知识元具有独立性、拓扑性和链接性。独立性是指每个知识元是一个独立的知识单位，都包含有知识点；拓扑性是指每个知识元是有其完整结构的，由知识元名称、知识元属性、知识元属性值组成，可以表示完整的知识内容；链接性是指知识元通过链接可以创造新知识，是知识标引的基础，也是知识元成为知识管理新纪元的关键。

由于知识资源的浩瀚和语义的复杂性，对知识元的分类与标引并非一件易事。知识元抽取可采用自上而下的分类方法，即知识元由六元组 CKIE 确定，知识类别指对知识体系按照学科标准分成若干领域，知识级别则按由浅入深的认知方式分成概念、公理、规则和方法 4 个层次。

在此，知识元标引起到知识元过滤和知识元链接的作用，从而为知识库的构建提供了有力保障。

5.2.1 定位知识元方向

识别向导信息，建立向导信息库：知识标引首先要从文献标题词入手，定位知识元方向。然后，通过文摘和关键词寻找知识元的向导信息，识别向导信息是知识元抽取的第一步。向导信息是标题、小标题、摘要、段首、段尾、结论、引文等其后有具体内容的特征词，如果在这一特征词后引导的段落和句子包含该特征词描述的知识元的内容，这一特征词就上升到了向导信息词的地位，将该特征词导入向导信息库；该特征词同时也成为知识元名称，将其同时导入知识元库。该特征词后引导的具体内容就是知识元内容。

知识元方向定位利用概念概括与划分的分析原理，通过建立面向对象的受控语言分类表和词表，并采用词位置及词频分析法完成，这些技术可借鉴利用主题标引研究成果；向导信息识别则利用链的思想。链是表示对象间物理与概念连接关系的一种实例，对象间的物理与概念连接关系则是链的抽象。如，“知识标引简单说就是以知识元为单位进行标引”，可表述如下：其中的关联词“是”表明了“知识标引”该特征词与其后的内容之间的关系，从而确定了“知识标引”向导信息的地位。同主题标引相比较，知识标引中识别向导信息的特征词就是主题标引中抽取的主题词。在主题标引中，主题词识别后抽取进入主题词库，直接用作标引用，主题词就是主题标引的内容；而知识标引的基本单位知识元的主要内容是特征词后的具体内容，不仅仅是知识元名称。

5.2.2 抽取特征句，获得知识元具体内容，完成知识标引，实现知识发现

以知识元名称为向导，在正文中抽取含有该名称的特征句若干，对每个句子中知识元名称进行词频统计，并按句子出现位置进行加权，筛选出其中的几个句子，作为该篇文献的知识元，用来对该篇文献进行标引。不同类型的文献及文献中不同位置的知识元表达的

连接关系不同。在创新性论文中抽取特征句时，文摘中的“本文的研究目的是……”、“本文发现……”、“本文对……作了改进”，正文中的“该方法称为……”、“该理论认为……”等；在包含数据型知识元的文献中抽取特征句时，含有时间、地点、数字等的文本内容；在包含事实型知识元的文献中抽取特征句时，“……就是指……”、“也就是说……”等是特征句选取的基本标识。

主要包括如下步骤：

(1)提取关键词。关键词，是指那些出现在文献的标题(篇名、章节名)以及摘要、正文中，对描述文献主题内容具有实质意义的词语，即对提示和描述文献主题内容来说是重要的、关键性的那些词语。判断关键词所在句是否包含知识元。

(2)句子分析。对提取出来的关键词所在句进行特征分类，为知识元标引做好准备工作。

(3)知识元标引。对做好特征分类的句子进行判断，看是否能成为一个知识元，完整地表达一个知识且不可再分。按照知识语义模型要求，获取知识元属性、所在文献的题目、知识元内容(即知识元所在特征句)、知识元的上下文等。

(4)知识元的生成。知识元可用 RDF/XML 文档表示，即在系统中可以根据文献、知识元类型等生成一个 RDF/XML 文档。

参 考 文 献

[1] 杨硕，崔蒙，赵英凯，刘明岭. 基于知识元的中医药信息知识标引[J]. 中国中医药信息杂志，2011，(8).

[2] 原小玲. 基于知识元的知识标引[J]. 图书馆学研究，2007，(6).

[3] 周和玉，杨元美. 文献数据库的知识标引研究[J]. 情报理论与实践，1998，(4).

[4] 姚小乐. LCSH，SKOS 和关联数据[J]. 现代图书情报技术，2009(3)：8-14.

[5] 温有奎. 知识标引与检索中的知识链研究[Z]. 西安：西安电子科技大学.

[6] 司莉. 基于知识构建的知识服务的实现[J]. 图书馆论坛，2009，(6).

[7] 曾新红. 中文知识组织系统形式化语义描述标准体系研究(一)——扩展 SKOS 实现传统受控词表的全描述[J]. 中国图书馆学报，2012，(2).

[8] 谈春梅，颜世伟，刘子牧. 网络专题知识组织知识元自动抽取系统的设计与实现[J]. 现代图书情报技术，2008，(3).

[9] 付蕾. 知识元标引系统的设计与实现[D]. 武汉：华中师范大学硕士学位论文，2009.

[10] 刘柏嵩. 基于 Web 的通用本体学习研究[D]. 杭州：浙江大学博士学位论文，2007.

[11] 中国图书馆分类法编辑委员会. 中国图书馆分类法(第四版)[M]. 北京：北京图书馆出版社，1999.

[12] 中国国家标准 GB/T 13745—2009《学科分类与代码》. 中华人民共和国质量监督检验检疫总局、中国国家标准化管理委员会于 2009-05-06 公布，2009-11-01 起实施.