产品分类本体构建的语义分析

钟毓灵

(大连海事大学 计算机科学与技术，大连 116026)

摘 要：形式概念分析（Formal Concept Analysis）在充分理解产品分类意义的基础上，指出了产品分类存在的五大问题，描述了产品分类的体系结构。针对这五大问题提出了产品分类本体构建语义分析模型，对每一要素进行了分析，重点突出了概念划分、类的处理、属性处理和语义细化，阐述了语义关系之间细微的语义差异。通过对产品分类进行语义分析，构建产品分类本体。

关键词：形式概念分析;形式背景;概念格

中图法分类号：TP18　　**文献标识码：**A

Semantic analysis of product classification ontology construction

ZHONG Yuling

(Computer science and technology, Dalian maritime university, Dalian 116026, China)

Abstract：After fully understanding the significance of the product classification，this paper points out 5 key issues of product classification，and then describes the ontology architecture of the product classification. For these issues，the paper proposes a se-mantic analysis model of the ontology architecture of product classification，and analyzes each element of the model，which high-lights the concept division，classification processing，property processing and semantic refinement. The paper expounds the subtle semantic differences between the semantic relations. Through the semantic analysis of product classification，the paper constructs the ontology of product classification.

Key words：product classification; architecture; ontology architecture; semantic analysis

0 引言

产品分类是现代物流的基础之一，目前正在蓬勃发展的电子商务以及新兴的物联网都离不开产品分类。产品信息的复杂性在于产品种类繁多，涉及各种行业，不同行业需要不同的产品属性，即使是同行业中的产品，由于类型不同，其属性描述也可能不同，而现在连最基本的产品分类也不统一。如胶鞋在生产领域属橡胶制品，与胶管、胶带分在一起;而在流通领域属鞋类，相差甚远。由于缺乏统一的产品分类标准及规范化的信息描述，人们犹如缺少一种交流语言，无法共享产品信息。

由于产品分类法采用的是等级体系分类方法，因此也具有等级分类体系的缺陷;归纳起来，存在如下五大问题:

1)类目线性排列与网状关系的矛盾。对于功能单一的产品，线性排列非常适用;而对于多功能、多用途的产品，线性排列的局限性是明显的。如按照分类原则，一个产品只能分在一个类目中，多功能、多用途的产品可能合适分到几个不同的类目。

2)语义表达能力有限。等级体系最能表达的是上下位类与同位类的关系。

3)概念的专指度受限。从等级分类体系原理上讲，概念的划分可以一直进行下去，使类目的末级达到很高的专指度。而现在一些主要的产品分类，如《全球产品分类》体系(GPC)为了编码的规范、简洁，采用了 4 个层级，这使得类目的末级显得混乱庞杂。

4)使用过于专业。分类法中隐含的联系和限定贯串于整个分类体系，对用户来说缺少易用性，不但有些类目本身的含义和类目间的逻辑联系对中小企业因雇不起这方面的专业人员而无法理解和掌握，其中有些技术甚至连专业人员也难以掌握。对于终端用户来说，检索需求千差万别，他们希望使用的对象有很好的易用性和可操作性。这也是他们中的很多，特别是那些中间件或配套件的生产者，还没有使用产品分类的原因之一。

5)周期长、更新慢。分类法主要以产品实体为分类对象，以逻辑划分为基础，兼顾产品属性。其产品范围划分相对稳定，体系严密而深细，更新周期比较长。

针对这五大问题，人们想了许多办法加以克服，如详细描述产品属性、建立产品数据字典等，对缓解困难有所帮助，但根本性的问题仍遗留至今。随着语义表示技术的进步，人们逐渐将目光聚焦在产品分类的语义化上，希望通过产品分类法的语义分析，构建产品分类本体解决问题。

目前国内外针对产品分类语义化的研究并不多见。在该领域研究中最为著名的是德国慕尼黑大学教授 M.Hepp，他系统性地指

出了构建和使用产品分类本体的六大困难以及描述产品的本体应涵盖的 6 个方面。K.Dongkyu 等以 UNSPSC 为例，从模型建立的不同方面，分析分类模式的种类，提出了语义分类模型，随后 H.Lee 等将本体模型应用到电子商务领域，利用扩展实体联系(EE R)和描述逻辑(DL)来构建电子目录本体，但并未建立起能够实现自动映射的工具，未在操作层面上实现本体的应用。由于上述的语义分析模型比较复杂，执行的效率不尽人意，D.Beneventano 等提出了一套半自动创建语义分析的方法，该方法可在不同的产品分类体系之间实现定义语义映射，并利用 UNSPSC、eCl@ss 及 eBay 的在线部分分类目录开展了该方法的实际操作展示。

此后，有关产品分类本体问题的研究逐步扩展开来。M.Hepp 进行了具有代表性的研究，构建了产品分类本体构建的本体，但是他仅仅是用 OWLLite 将 eCl@ss 分类系统书写了一遍，缺少对原有产品分类体系的再加工与知识提炼，没有对概念之间的语义关系做深入分析，对原产品分类体系中正确与错误的语义关系都予以承认，并保留在新建的本体中，造成所构建的产品本体在推理和复用层面都表现欠佳，无法满足电子商务大规模的数据交换。

本研究就是在 M.Hepp 等人工作的基础上，提出一套产品分类本体构建的语义分析方法，试图对“五大问题”的解决有所进展。

1 产品分类的体系结构

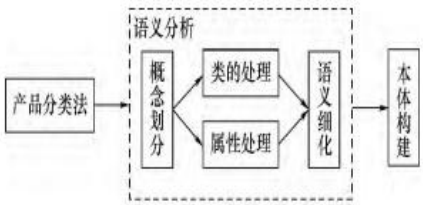
产品分类体系拓扑结构普遍采用了简单的树形结构，目前电子商务中使用最广泛的产品分类法是 GPC(全球产品分类法)，而发展最迅捷的分类法是 eCl@ss(先进分类系统)。本文选取 GPC 和 eCl@ss 进行介绍。GPC 和 eCl@ss 都采用四级类目结构。GPC 的 2013 标 27 个专业，包括 584 个主类，4982 个分类，27200 个支 ss 每层用两位阿拉伯数字表示其类目编码，从第 1 层逐层分到第 4 层，类目编码由左往右连续排列，共 8 位。GPC 的第 1 层表示产品的大类，为产品隶属的行业;第 2 层表示产品的中类，为产品隶属小行业;第 3 层表示产品的小类，为产品隶属的族;第 4 层表示产品的细类，具体的产品品种即基础产品类别。eCl@ss 的第 1 层为产品的专业范围;第 2 层为产品的主类;第 3 层为产品的分类;第 4 层为产品的支类，具体的产品品种即基础产品类别。GPC 和 eCl@ss 都在第 4 层定义了属性。GPC 属性的描述字段由属性代码、属性名称、属性定义和属性值组成;属性值的描述字段由属性值代码、属性

值定义和属性值内容组成;eCl@ss 的属性又分为基础属性(BSP)和专有属性(SSP)，每条属性的描述字段增加了属性名简称、数据格式和计量单位，其余与 GPC 完全相同。两者之间的结构关系见图 1，虚线框内容为 eCl@ss 所特有，其余为两者共有。

2 产品分类本体构建语义分析模型

语义分析最早出现在哲学领域，按其发展历程可分为 4 个阶段:基于句法的语义分析、基于相似度计算的语义分析、基于本体构建的语义分析和基于本体推理的语义分析。前两种方法仍停留在对自然语言处理，处理的是句子的内在语法结构和词频的统计特征;后两种方法与本体密切相关，处理的是概念之间复杂的语义关系。一个具有强大功能的产品分类本体就需要包含尽可能多的语义关系。

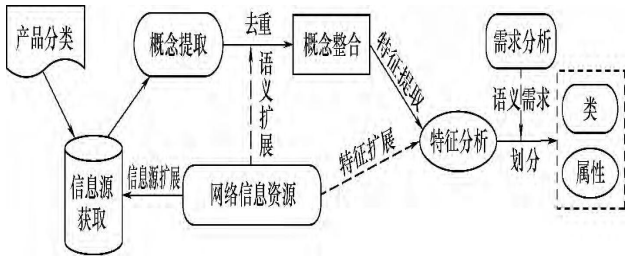
因此，本文设计了产品分类本体构建的语义分析模型进行产品分类本体构建，着重分析产品分类本体的属性、类和语义关系的处理(见图 1)。



2.1 概念划分

概念是本体的重要组成部分，概念包含属性和类，概念划分就是将获取的概念进行分类，依照概念的特征(包括外延特征和内涵特征)通过语义分析对其进行属性或类的划分。外延特征指的是将概念各组成部分进行有机组合，从概念的整体结构角度出发，考虑各组成部分之间的联系;而内涵特征指从概念自身所有的内涵本质出发，主要考虑各组成部分各自的特征结构，而忽视它们之间的联系。

该模块主要包括产品分类信息源获取、概念提取、概念整合、概念划分和语义需求等部分。(见图 2)



1）产品分类信息源获取。指对产品分类概念信息的获取，是进行产品分类本体构建的基础。一般而言，信息源的获取渠道很多，为了更好地体现不同产品分类法自身的特性，本研究的信息源获取主要来自于产品分类相关文档及已有产品分类本体，网络信息资源只作为其必要扩展和补充。

2）概念提取。指对信息源获取的数据进行处理，采用一定方法进行概念提取。目前本体概念获取的主要方法有:从叙词表获取概念、从专业词典和专业书籍的术语中获取本体概念、从原有的本体中获取概念、通过搜索引擎获取概念、从已有文档概念抽取获取概念和从领域论文关键词中获取概念等方法。本研究主要采用从搜索引擎获取概念和已有文档概念抽取概念两种方法。

3)概念整合。指对获取的概念进行去重，并利用网络资源进行扩展是对概念扩展和整理的过程。

4)特征分析。指对概念的内涵特征和外延特征进行分析，包括特征提取和特征扩展。特征提取指的是对整合后的概念，在现有资料的基础上进行特征分析;特征扩展则指的是利用网络信息资源，对已获取概念的特征进行扩展分析。

5)语义需求。指结合应用场景，依据特征分析结果，将整合的概念划分为属性或类。

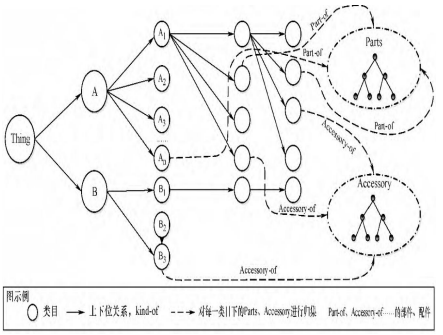
2.2 类的处理

类是产品分类本体的核心，用来描述产

品分类的念。类的处理指对获取的类目进行细化和调整，是语义进一步丰富的过程包括:确定类的上下位关系、类的层级数、同级类目数、类间关系和等价类。

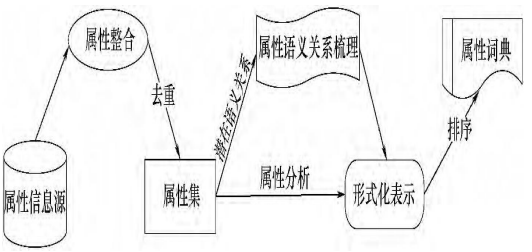
对于类的上下位关系，其处理有 3 种可选方案，分别是自上而下、自底向上和由内向外策略;本模型选用自上而下的方法进行类的处理，这种策略首先定义产品分类类领域中最上层的类，然后对类依照概念划分中的特征分析结果，进行特征化，将类按照其特征逐一细化，以推得到类的上下位关系。于类目的层级数和同级类目数，其处理结合概念划分段具体的语义需求，对特征丰富、语义需求高的类目要增加层级数和同级类目数;对特征欠缺、语义需求低的类目要减少层级数和同级类目数。对于类间关系，依照本体的“五元说”，通过公理实现。对于等价类，采用 Nan-da 等提出的基于产品簇的本体构建方法(PFODM)来进行判断;其基本思路是从产品设计的角度出发，利用形式概念分析方法(FCA)，分析产品各构件，通过各构件之间的相关性，判断类与类的等价关系。

类的处理解决了类目线性排列和网状关系的矛盾。类目线性排列与网状关系矛盾的根本原因是产品的多功能、多用途导致的，一般的分类方法往往只关注了产品的某一个或某几个方面的功能，而忽视了产品的其他功能，若不同类目采用的分类标准不统一，则会导致同一个产品可以被好几个类目所共有。解决这一问题的方法之一，就是利用定义一些特殊功能的类进行解决。如在类的处理过程中，考虑到产品分类的特殊性，本模型在定义最上层类目时，定义了 Thing 的直接子类 Parts 和 Accessory，将产品分类中有关部件和配件的类目都放到对应类目下。在 Parts 和 Accessory 中又按照类的上下位处理方法，对其进行分类;通过 Part-of 和 Accessory-of 关系将其与相对应的类进行关联(见图 4)。



2.3 属性处理

属性处理指对属性集进行语义分析，构建属性词典。构建属性词典实现了对产品分类语义的进一步丰富。主要包括:对获取的属性信息源的形式化表达;属性与属性内在语义关系的分析;属性的分类和属性词典的构建。该模块主要包括属性整合、属性语义关系梳理、形式化表达等部分(见图 5)。



1)属性整合。指对概念划分得到的属性进行整合，去重得到属性集的过程。

2)属性语义关系梳理。指分析属性集各属性间的潜在语义关系，决定属性与属性之间的内部层次结构;同时对属性集中每个属性进行属性分析，包括属性的属性值、属性名、属性编码和属性定义等。

3)形式化表示。指将属性语义关系梳理后的属性集，利用 OWLDL 语言进行属性形式化表达，并进行排序得到属性词典。类和属性处理、语义细化解决了概念专指度受限的问题。概念专指度的问题是由现有的分类体系自身结构缺陷所导致的;从等级分类体系原理上讲，类目的划分可以一直进行下

去，使类目的末级达到很高的专指度。但在现实中，现在一些主要的产品分类，如上文介绍的 eCl@ss 为了编码的规范、简洁，仅采用了 4 个层级，并且属性也没有进行层级划分，仅划分为基础属性(BSP)和特有属性(SSP)两类。这使得类目的末级显得庞杂，甚至出现部分四级类目根本不是对上一级类目进行细分的情况，仅为了分类体系结构的完整，凑整凑出了四级类目。解决这一问题的方法就是利用本模型类的处理、属性处理和语义细化。类的处理解决了产品分类体系四级类目的限制;属性处理明确了属性之间的层级关系;语义细化确定了类与属性的从属关系。从而丰富了概念，解决了产品分类概念专指度有限的问题。

2.4 语义细化

在获取到的单值形式背景的基础上做顺序的调整，找到属性继承的父子关系，例如 7 可由对 6 的全部属性继承的基础上添加自身属性 d 得到；8 可由对 6 的全部属性继承的基础上添加自身属性 b 得到。通常情况下，为方便查找，从上倒下按属性的多少进行排列。表 4 所示为最后形成的单值形式背景。

语义细化指的是考虑类、属性和实例相互之间的语义关系，通过语义分析对它们间的关系进行语义细化。按照语义细化的对象来分，语义关系主要包括类—类、类—属性、类—实例、属性—属性、属性—实例、实例—实例这六大类(见表 1)现有的产品分类本体是对产品分类法原封不动的形式化表达，缺乏对产品分类体系内在语义的分析，类与类之间的语义关系笼统。本研究提出了语义细化方法，为了比较细微的语义差异，引入了集合论和描述逻辑语法对它们之间的区别进行语义刻画。表 1 中某些关系的区别是显而易见的，在此不加赘述，本研究主要讨论表 1 中经常使用和易于混淆的部分。

对象	关系	表示	含义
类—类	种属	Kind-of	价，都是对同一事物的描述，只是概念范围的大小
	部分与整体	Part-of	表示类与类间整体与部分的关系
	类与配件	Accessory-of	表示一个类是另一个类的配件
	类与用料	Materialsof	表示一个类是另一个类的用料
	类组合	Make-up	表示一个类由其他几个类组合而成
类—属性	从属(抽象)	Attribute-of	表示某一个类拥有属性
类—实例	抽象与具体	Instance-of	抽象与具体，类似面向对象中对象与类关系
属性—属性	上下位关系	SubAttribute-of	表示属性之间的上下位层级关系
属性—实例	从属(具体)	Attribute-Instance	表示某一实例具有的属性
实例—实例	具体与具体	Instance-Instance	表示实例与实例之间特有的语义关系，主要包括相交和不相交关系

1)Part-of 与 Kind-of 的语义细化。Part-of 和 Kind-of 描述的都是类与类之间语义关系，在本体描述语言 OWL 中都是 SubClass 关系。Part-of 更多强调的是部分与整体的关系，而 Kind-of 强调的是类与类之间上下层级间的种属关系。在 Part-of 中一个类必是另外一个类所描述事物的一个组成部分;而 Kind-of 中一个类是另外一个类的一种，是同一事物在粒度上的差异。例如:计算机按用途分为超级计算机、工业控制计算机、网络计算机、个人计算机、嵌入式计算机;计算机由控制器、运算器、存储器、输入和输出设备组成。

2）Part-of 与 Accessory-of 的语义细化。Part-of 与 Ac-cessory-of 都是描述的类—类关系，在本体描述语言 OWL 中均是 SubClass 关系。Part-of 描述的类—类关系，一个类是另一个类的部分，且对作为部分的类没有特殊要求;Accessory-of 更强调通用性，即作为部分的类必须被其他类所共有。通用性强的部分与整体关系被定义为 Accesso-ry-of，通用性弱的 SubClass 关系则应被定义为 Part-of。

3）Instance-of 与 Kind-of 的语义细化。

Instance-of 与 Kind-of 两者有着本质的区别，Instance-of 描述的是类—实例的关系，而 Kind-of 描述的是类与类的关系。可用集合论的表达方式对两者进行区分，设类的集合 A={A1，A2，…，An}，则元素 A1A，关系刻画的就是 Instance-of;而对于类的子集{A1，A2，A3}A，关系刻画的就是 Kind-of。所以在进行语义细化时，对于类与实例的关系要用 Instance-of 描述，而对于类与类的关系则要用 Kind-of 进行描述。

语义细化解决了语义表达能力有限。语义表达能力有限是由等级体系只能表达上下位类与同位类的关系而造成的，上下位关系只用 SubClass 父子类关系进行表达，而在本体中，SubClass 体现的是种属关系(Kind-of);而在产品分类法中，SubClass 可以体现种属关系(Kind-of)、部分与整体(Part-of)、类与配件(Accessory-of)和类与用料(Materials-of)等关系;并且在类—属性、类—实例、属性—属性、属性—实例、实例—实例间，也存在类似关系。

3 产品分类本体构建语义分析的应用

本研究依据上述模型，构建了 GPC 的的本体。概念划分阶段共整理出 598 个 GPC 概念和 2454 个 eCl@ss 概念，其中 598 个 GPC 概念，均被划分为类目;而 2454 个 eCl@ss 概念，有 1316 个概念被划分为属性，其余的 1138 个概念均被划分为类。类处理和属性处理阶段，分别根据上述分类结果进行类和属性的梳理，GPC 是采用其现成的 GDD 属性词典;eCl@ss 是对获取的 1316 个概念按照其语义分析模型分析后，对属性进行层次分类，构建属性词典。在语义细化阶段，构建了 Part-of、Kind-of 和 Make-up 等语义关系，并对其进行了定义，部分表达的形式化代码如下所示:

```
<owl:objectProperty:ID="make-up">
<rdfs:domain>
<owl:Class>
<owl:unionOf:parseType="Collection">
<owl:Classrdf:about="#Scanner_head">
<owl:Classrdf:about="CCD"/>
<owl:Classrdf:about="#Scan_rod_drive_motor"/>
<owl:Classrdf:about="Glass_countertops"/>
<owl:Classrdf:about="#Control_buttoris_circuit_board"/>
<owl:Classabout="#Other_scanner_accessories"/>
</owl:unionOf>
</owl:Class>
</rdfs:domain>
<rdfs:commentrdf:datatype="http://www.w3.org.201/XMLSchemaJHJstring">
表示一个概念由其他及格概念组合而成
</rdfs:comment>
<rdfs:rangerdf:resource="JHJHand-held_scanner"/>
</owl:ObjectProperty>
```

4 结束语

本文通过类的处理解决了类目线性排列和网状关系的矛盾;通过类和属性处理、语义细化解决了概念专指度受限;通过语义细化解决了语义表达能力有限;通过概念划分已从源头上解决使用过于专业的问题，但真正的解决需要产品分类本体查询系统来实现;周期长、更新慢的问题仅通过产品分类本体构建是无法得到有效解决的。

参考文献

[1] 黄映辉.智能信息处理课件.大连海事大学,2014.

[2] 干特尔，威尔，马垣.形式概念分析.科学出版社，2007.

[3] 曲开社.偏序集、包含度与形式概念分析.计算机学报，2006，29(2):219-226.