从文本中获取植物知识方法的研究*)

罗 贝[1] 吴 洁[1] 曹存根[2] 邵志清[1]

(华东理工大学计算机科学与技术专业 上海 200237)[1] (中国科学院计算技术研究所 北京 100080)[2]

摘 要 知识获取一直是人工智能中的一个关键问题。当前,知识的文本挖掘(KAT)已经成为计算机领域的一个重要的研究课题。本文中,给出了基于植物本体的从海量网页文本库中自动获取植物领域知识的方法。该方法包括两个部分,一是植物本体(Botanical Ontology),它是顾芳博士等建立的生物本体的扩展。第二部分是以植物本体为基础,在网络文本库中进行文本挖掘(Text Mining),自动获取植物知识。实验证明,基于本体的文本挖掘是一种有效的知识获取方法。

关键词 植物本体,植物知识,基于本体的知识获取与分析

Botanical Knowledge Acquisition from Text

LUO Bei[1] WU Jie[1] CAO Cun-Gen[2] SHAO Zhi-Qin[1]

(Computer Software and Theroy, East China University of Science and Technology, Shanghai 200237)[1]

(Institute of Computing Science, Chinese Academy of Sciences, Beijing 100080)[2]

Abstract Knowledge acquisition is the bottleneck of artificial intelligence. Knowledge acquisition from text (KAT) is one of the most researched topics in the current computer domain. In this paper, we present an Ontology-based method for acquiring the knowledge of botany from a large corpus of Web pages. The method consists of two components. One is an Ontology of botany. This Ontology is an extension of Gu et al.'s Ontology, and is an important base for the second component—the knowledge acquisition from Web pages. our knowledge acquisition method is Ontology-driven in the sense that the content in the Ontology is an efficient guide for what knowledge to extract to fill in the frames in the Ontology.

Keywords Ontology of botany, Botanical knowledge, Text mining, Ontology-based knowledge acquisition, Knowledge analysis

1 引言

如何获取、处理和应用知识已经成为当前很重要的一个研究课题[1]。现阶段的大规模知识处理都致力于各学科专业领域的知识,并从知识的角度来探讨智能的实现方式。所以,建立各领域的知识库对研究人工智能有着十分重要的理论和现实意义。本文的研究重点是如何从 Web 网页中获取植物学知识,并且建立相应的植物学知识库。

植物学是与人类密切相关的科学,植物知识也当然成为人类最重要的知识,因此世界各国都在进行植物学知识工程的研究。建立一个良好的、完备的植物学知识库具有极其重要的意义,它将为植物学专家系统,植物学信息检索,植物教育系统,自然语言理解等领域提供智能基础。中科院现在正在建立一个大型的国家知识基础设施,我们的植物学知识库是国家知识基础设施的一个子集。植物知识库的建立对 NKI 的建立,人工智能的研究都有极其深远的意义。

BKB(Botanical Knowledge Base)是美国 Texas 大学花了近 10 年的时间才建立起来的一个植物知识库,基本采用手工的方法[2,3]。与 BKB 不同,本文将介绍一种较为自动的、基于植物本体的文本知识获取方法。

本体是一种关于领域概念的形式刻画[1,4],通常是由概念、公理和关系所组成的高级描述。概念是用来对知识库进行组织的上层部分,公理用来从逻辑上定义知识之间的特有联系,而关系则是概念与概念之间的联系。本体类似于一个辞典或词表,但是又比它们更详细,其结构可以使机器直接读取。从知识共享的角度来说,本体作为一种概念化的说明,采用框架系统对客观存在的概念和关系进行描述。它是通用意义上的"概念定义集",是关于"种类"(kind)和"关系"的词汇表。这种词汇表,是在各种事务代理之间交换意见时所用到的共同语言。由于本体工程到目前为止仍处于相对不成熟的阶段,每一个工程拥有自己独立的方法[1]。目前应用比较多的如美国 D. Lenat 教授领导小组正在研制的大型常识知识库系统 Cyc,Princeton 大学研制的语言知识库 WordNet,国内陆汝钤院士等研发的常识知识系统,以及现在正在进行中的大规模知识获取研究[4,5]。我们引入植物本体的概念来进行植物知识的获取,并根据植物本体来建立知识获取的方法和验证分析系统,将所研究的方法放到文本库中使用,最终建立 NKI 的植物知识库。实践证明,这是一套行之有效的方法。

本文第 2 节介绍植物本体,包括植物的体系框架,本体描述语言,植物本体的属性和关系等,以及用来验证的公理系统。第 3 节介绍植物概念知识的获取方法。在第 4 节介绍实验结果和分析。最后进行总结。

2 植物本体概述

2.1 植物本体体系

植物学是一门研究植物形态解剖、生长发育、生理生态、系统进化、分类以及与人类的关系的综合性科学,是生物学的分支学科[6]。我们的植物本体也是基于植物学体系而建立。

*)本文工作得到自然科学基金的资助(#60273019,#60373075,#60496326)和科技部重大基础项目基金(#2002DEA30036)的资助。罗 贝 硕士研究生,研究方向为计算机软件与理论。吴 洁 硕士研究生,研究方向为计算机软件与理论。曹存根 博士生导师,主要研究方向:人工智能。邵志清 博士生导师,主要研究方向:Generic 程序设计技术,网络信息服务技术,软件开发与验证方法。

自然界的植物，种类繁多，形态各异，充分表现出植物的多样性。到现在为止，已知的植物约有 50 多万种，其形态、结构、

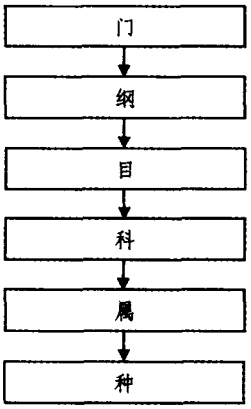

图 1　植物分类结构

生态习性以及对环境的适应性各不相同，千差万别。世界上如此多的植物，我们如何进行植物本体体系建立的呢？植物本体体系的建立，必须经过一定的步骤。首先需要将性质相近的植物进行分门别类，然后寻找各类群的相互关系（即亲缘关系），再根据其关系的密切程度，加以排列，这样就可以从系统中看出整个植物界，或者是某一门类植物的特性和发展[7,8]。而基于植物不同的共同点，就会有不同的分类法，从而描述植物的角度也会有不同。根据植物在进化中所形成的特点，通常将地球上的植物分成两大类：高等植物和低等植物；若根据植物的用途、形态和习性等方面可以分为粮食植物、纤维植物、油料植物、药用植物等；而根据植物的亲缘关系和演化规律来分类时，在植物界下又分门，可分为 15 个门：蓝藻门、红藻门、绿藻门、裸藻门、甲藻门、轮藻门、金藻门、褐藻门、真菌门、细菌门、粘菌门、地衣门、苔藓植物门、蕨类植物门和种子植物门，以下又设纲、目、科、属和种[6,8]。结构见图 1。例如陆地棉这种植物在分类上就是：植物（界），种子植物（门），被子植物亚门（亚门），双子叶植物纲（纲），锦葵（目），锦葵科（科），棉属（属），陆地棉（种）。

由于植物在长期发展的历史过程中许多植物的类群已经灭亡，所遗留的痕迹只有在地层的化石中存在，而化石材料又残缺不全，因此，在编制植物分类系统时存在很多困难。在设计植物本体的过程中，我们基本按照图 1 中的分类结构，将整个植物界中的植物按照其性质归纳成 15 个门，而在每个门中又包括多种植物，例如，被子植物门中包含有 20 多万种植物。

因此，分类仅分到门是不够的，门只不过是植物分类的最大单位，包含在同一个门的植物还可以继续分下去，分成许多单位，直到不易再分为止。最小单位种是生物分类的基本单位，它是具有一定的自然分布区和一定的生理，形态特征的生物类群。同一种中的各个个体具有相同的遗传性状，而且彼此杂交可以产生后代。但与另一个种的个体杂交，在一般情况下，则不能够产生后代[8]。所以种是我们分类的最小单位。基于这种分类结构，我们通过对植物专业知识的分析，以及与植物专家讨论，建立了植物的本体体系，具体结构见图 2。

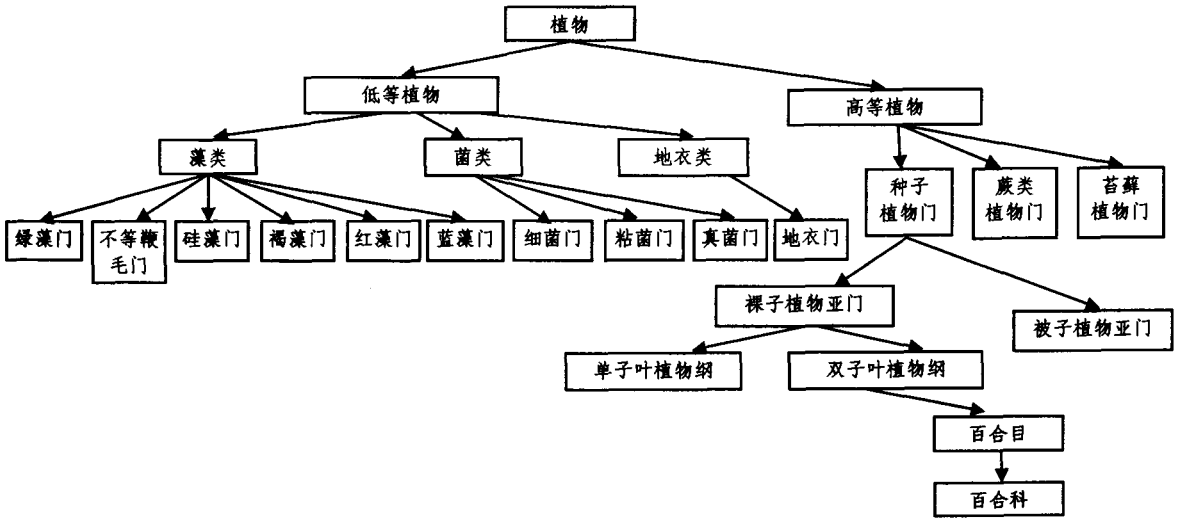

图 2　植物本体体系（部分）

从图 2 可以看出，把植物分为两个大类：高等植物和低等植物。前者主要有三个门类，分别是种子植物，蕨类植物，和苔藓植物。而低等植物也有三大类，分别是藻类，菌类，和地衣类，每个类有多个门。门下面为纲、目、科、属、种的分类。有的会细分为亚纲、总纲、次亚纲等。该图大致反映了整个植物界各大门类之间的系统演化上的相互关系。低等植物各门，在进化上处于比较低级的地位，这不仅从形态、结构和生殖方式上反映出来，而且也反映在它们的生态特性中。种子植物，特别是被子植物，处于进化的最高阶段。它们与处于最低阶段的藻菌门植物很少有相似之处。处于中间阶段的苔藓植物与蕨类植物，虽然被人习惯地划入高等植物的范畴之中，但其本身既具有高等植物的特征，也具有低等植物的特征。这正反映了它们的过渡性质[8]。

很明显，图 2 描述的本体体系就是基于我们的基本植物分类机制，从父节点到子节点，层层细化形成了如上图的树状结构，构成了我们的植物本体体系。可以看出，这样的体系一般不会产生知识交叉现象。即，一种植物若属于一个门，或纲，那它不可能同时属于另一个门，或另一个纲。所以，用这样的体系结构在对植物进行描述的时候就比较清晰明了，避免了产生信息冗余的情况。

2.2　本体描述语言

我们在描述植物学本体时使用的是基于框架的本体语言。该语言是一种框架语言，主要的文法如下所示[5,9]。

〈本体描述〉::=〈类定义〉{〈类定义〉}{〈公理定义〉}
〈类定义〉::=〈定义词〉〈类名〉[〈相关类〉]
{
　{〈类知识描述〉}

{〈槽定义〉}
{〈类内公理定义〉}
}
〈定义词〉::=defcategory|defprivatecategory

其中,Defcategory 定义的是一般的类,能为其他类所继承。而 defprivatecategory 定义的是内部类,它的属性和关系不能为外界所继承。

类定义的核心内容为槽及类公理。槽代表了本体的属性及与概念相关的知识,是我们在植物学本体中的描述的重要组成部分。其描述文法如下:

〈槽定义〉::=〈槽类型〉:〈槽名〉
:类型〈槽值类型〉
[:值域〈值域〉]
[:单位〈槽值的单位〉]
[:同义词〈同义词序列〉]
[:近义词〈近义词序列〉]
[:侧面〈用户定义的必要侧面序列〉]
[:注释〈槽的非形式化说明〉]

槽分为 4 个类型.

1. 属性槽(简称属性):属性为名词。属性槽又分为布尔属性槽和非布尔属性槽。布尔属性槽对应于一元谓词。

2. 关系槽(简称关系):关系为动词,约束了概念之间的联系。

3. 属关槽(简称属关):既可以为名词又可以为关系的槽。例如:"又称"、"简称"等等。

4. 方法槽(简称方法)。

槽值的类型比较复杂,可分为简单类型和复合类型。其中,简单类型有:整数、数量、实数、分数、比例数、字符串、时间、日期,等。复合类型有:整数数组、数量数组、实数数组、分数数组、比例数数组、字符串数组、时间数组、日期数组,等。

相关类:一种是继承关系的相关类,一种是实现关系的相关类。

〈相关类〉::=继承〈继承类序列〉[;实现〈实现类序列〉]|
实现〈实现类序列〉[;继承〈继承类序列〉]
〈继承类序列〉::=〈概念类〉{,和〈概念类〉}
〈实现类序列〉::=〈聚类属性类〉{,和〈聚类属性类〉}

一个类 C1 继承另一个类 C2 表示三个含义。1)C1 可以使用 C2 中的属性、关系、槽和方法;2)C1 遵循 C2 中的所有公理;3)若 P 是 C1 中的概念,则 P 必然是 C2 中的概念。例如:上面给出的植物本体体系中,植物根类继承植物部分类,那么杨树的根作为植物根的概念,同时必然也是植物部分类中的概念[9]。

2.3 植物本体的属性和关系

2.3.1 *植物本体的属性* 描述一种植物对象,必然会从植物的性状特征和生态习性来加以描述。所以,我们也将从这两方面来定义本体中植物的属性类。植物的很多属性是生物界共用的,如门、纲、目、科、属、种的分类,还有如生殖方式,营养方式等[10]。它们都是继承生物类的属性。所以我们在生物类中定义了生物界的共通属性类,即:生物分类属性、生物进化属性、生物物种状况属性、生物分布属性、生物构成属性[2]。如生物分类属性中类中就有门、纲、目、科、属、种等的属性描述。植物和其他生物共通的属性就通过生物共通属性类来继承。这样做的好处是可以实现属性的复用,在描述动物本体的时候,也可以继承生物类的这些属性。

对于植物部分所特有的属性,我们从专业的角度对其划分为了植物细胞、植物根、植物茎、植物叶、植物花、植物果实、植物种子等几部分。在这些分类的基础上,对各种植物特征属性进行归类,抽象出植物细胞类、植物根类、植物茎类、植物叶类、植物花类、植物果实类、植物种子类等几个属性类。图 3 给出了植物茎属性类的一部分。

```
defcategory 植物茎
{
    是子类: 植物部分
    //植物的茎的其他属性从植物部分中继承
    属性: 类型
        : 类型 字符串
        : 值域 营养茎, 和 生殖茎, 和 直立茎, 和 匍匐茎,
          和 缠绕茎, 和 攀援茎
    属性: 质地
        : 类型 字符串
        : 同义词 茎干质地
        : 值域 草本, 和 藤本, 和 木本
    属性: 形态
        : 类型 字符串
        : 值域 直立, 和 匍匐, 和 缠绕, 和 攀援
        : 注释 "植物部分中已定义'形态'属性, 此处定义为
          说明其值域"
    属性: 粗细度
        : 类型 字符串
        : 值域 粗, 和 细, 和 较细
    属性: 易折性
        : 类型 字符串
        : 值域 容易, 和 不容易
    属性: 易断性
        : 类型 字符串
        : 值域 容易, 和 不容易
    属性: 分枝方式
        : 类型 字符串
        : 值域 不分枝, 和 分枝, 和 假单轴分枝, 和 单轴
          分枝, 和 二歧式分枝, 和 合轴分枝, 和 分蘖
}
```

图 3 植物茎属性类(部分)

从图 3 可以看出,我们对植物的茎这个属性类定义了很多属性,如茎的类型,质地等。从理论上说,属于茎的特征属性的,都会作为一个属性添加进去。这样就可以全面、精确、详细的描述一个植物各方面的特征。

2.3.2 *植物本体的关系* 仅有植物的属性类还不能够完整地描述植物的生态特性。如:一个纲必然要属于一个门,一个种必然要属于一个属,植物的某个部分可能是由另一个部分进化而来,或者退化而来等等。这些属性之间的关系也是植物特性的一个很重要的组成部分。所以,除了本体中的属性,还定义了一些关系,用来描述概念与概念之间的联系。这些关系与其属性结合在一起,对知识的获取,以及后面的公理的构建都有很重要的作用。我们从植物界的特征和规律方面,对植物中的属性之间的关系进行了分析,总结了属性之间的联系和约束,构成了我们的植物本体中的关系体系。

下面给出了分类系统中的部分关系表示的一部分。

```
关系: 界属
        : 类型 字符串
        : 侧面 分类系统
        : 注释 生物所属的界, 如属于 "动物界", "植物界"
关系: 门属
        : 类型 字符串
        : 侧面 分类系统
        : 注释 生物所属的门, 如: 属于 "种子植物门", "绿
          藻门" 等
关系: 亚门属
        : 类型 字符串
        : 侧面 分类系统
        : 注释 生物所属的亚门, 如: 属于 "裸子植物亚门",
          "被子植物亚门" 等
关系: 有部分
        : 类型 字符串数组
        : 同义词 有结构
        : 注释 "列举出生物部分, 可能是不完全的列举.
          cf '组成划分'"
关系: 进化为
        : 类型 字符串数组
        : 同义词 进化成
        : 注释 "生物直接进化成的新物种"
```

图 4 植物属性关系(部分)

2.4 植物本体的设计

植物本体的属性和关系定义好以后，我们就基本上完成了对植物本体类的定义。其结构框架见图5。我们将对植物的描述分为植物和植物部分两类。植物类主要是关于植物体本身生态性状上的特征，而植物的某一部分，如根，茎，叶等各自都有其独立的描述语言，所以我们把它们作为另外一个部分来描述。植物类和植物部分类都继承于相应的生物类和生物部分类。图5中的箭头代表上下继承关系。最上层的生物类和生物部分类属于抽象类，定义了如前面所述的生物界植物和动物共通的属性和关系，同时也用来实现一些属性类。

在设计植物领域本体的属性与关系时，原则上尽量使用专业化的描述。这样可以使植物本体的框架完整化，专业化。有利于不同层次的知识检索。

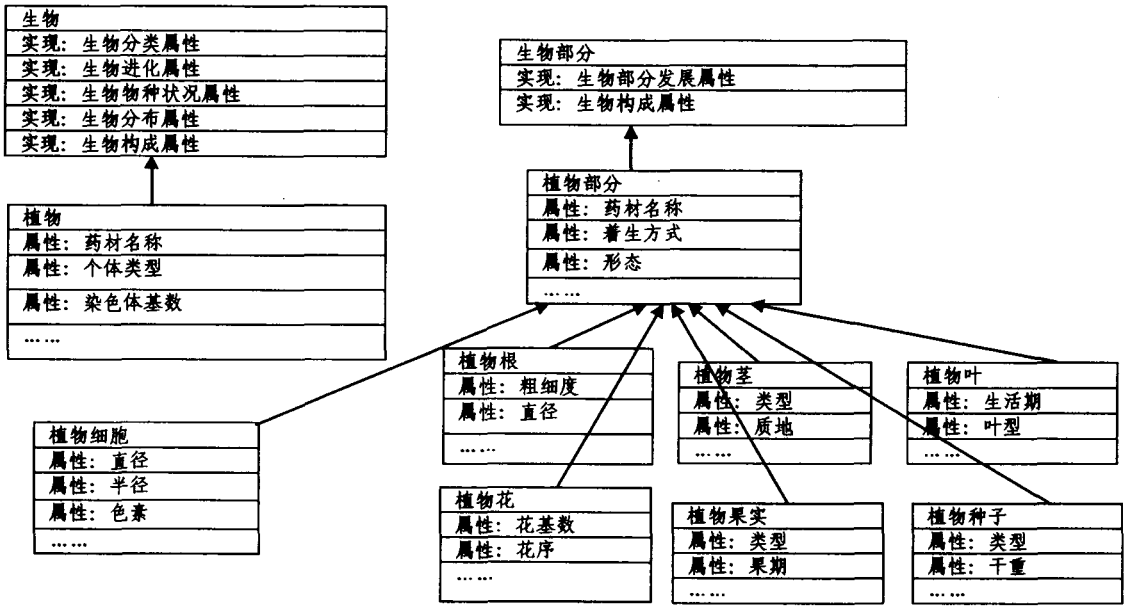

图5　植物本体类框架图

在设计本体中的槽时，需要根据上面提到的分类和结构，参考植物方面权威的书籍，并与植物方面的专家进行讨论协商，尽量使我们设计的槽能尽可能完备的描述整个植物界的所有属性。并且，这其中不能有重复的，冗余的结构。这一方面是本体设计的难点，一方面也是本体的特点所在。正如中科院的曹存根教授所提出的，要“言无不尽”。

对于本体中槽的值域，将其分为两类，一类为可收敛的，一类为不可收敛的。对于一些属性如：“营养方式”，其可能的值只有：“自养”、“异养”和“厌氧”[8]，诸如此类的值域可以完备其可能的值，我们称之为可收敛的。相反，对于如：“分布地”等发散的，不可能完备其值域的属性，则称为不可收敛的。原则上，对可收敛的属性的所有值域要进行完备化。可以从书籍，文本库中整理获取这些知识[12~14]。

通过对植物本体类框架的细化和填充，就基本完成了植物本体的设计，我们就可以利用本体，从全方位完整，规范的描述一个植物界的任何一个实体。这时候，就可以把一个植物概念看成是本体中某个类的实例。如可以把植物红皮云杉作为植物类的一个实例[7]。具体如下：

```
defframe   红皮云杉: 植物类
{
      拉丁名: Picea koraiensis Nakai
      同物异名: 红皮臭，虎皮松
      科属: 松科
      属属: 云杉属
      质地: 乔木
      高度: 30米
      花期: 4月
      果熟期: 10月
      性别分化: 雌雄同株
      原产地: 东北亚地区
      中国分布: 我国东北小兴安岭等地有分布；山东各地广为栽培
      繁殖方式: 播种繁殖
      实用价值: 因其耐阴性强，可用于建筑的背阴面
}
```

2.5　植物公理

基于我们的植物本体，我们从文本库获取出大量的植物知识。由于文本库中的文本知识广而杂，信息源本身也不能保证其绝对的正确性，所以需要有套分析系统来对获取出来的知识进行分析和校验。所以植物公理是植物本体的一个很重要的组成部分。我们根据植物界的专业知识和自然常识整理出了一套公理系统，可以通过这套公理系统，利用相关的规则，对知识进行分析。我们这里所说的公理是植物知识之上更高层的知识，是关于植物知识的知识，是植物学元知识，具有很高的抽象性。用来表示诸如概念与概念，关系与概念之间的内在联系[9]。

2.5.1　*植物公理体系结构*　从植物本体的架构可以看出，植物本体的诸如：门、纲、目、科、属、种的这种上下位结构性很强，也很重要，它构成了整个植物本体的大框架。由于其重要性和特殊性，将这部分植物结构（Backbone Structure，简称BS）公理单独作为一个块进行分析。对于非植物结构公理，如植物部分，植物生命特性等作为另一块。于是我们的植物公理系统分为了两部分：植物结构公理，非植物结构公理。这两部分会使用不同的公理表示方式。我们的公理体系包括了植物领域的公理，以及相关的领域知识和由公理推导出来的规则。

2.5.2　*植物公理表示方法*　首先介绍植物结构公理表示方法[2]。植物界的结构框架元素由：界、门、纲、目、科、属、种构成。其关系见图2。将植物结构（BS）用一个四元组表示：$BS=\langle SR,V,E,\alpha\rangle$，其中：

$SR=\{$属于界，属于门，属于纲，属于目，属于科，属于属，属于种$\}$

$V=\{C|C$是一个门，纲，目，科，属，种$\}$

$E=\{\langle v1,v2\rangle \mid v1\in V,v2\in V\}$

α 是一个 $SR\rightarrow E$ 的映射。对于所有 $p\in SR$，$\alpha(p)=\{\langle v1,v2\rangle$|当 $p(v1,v2)$在植物界中成立$\}$结构类＝def{界类，门类，纲类，目类，科类，属类，种类}

植物结构公理表示模式如下：

所有 X,Y,Z:结构类，所有 $P1:SR$，所有 $P2:SR$，$P1(Y,Z)\wedge P2(Z,Y)\rightarrow P2(Z,X)$

表 1　植物公理中用到的谓词(部分)

谓词与函数	含义
等于(X,Y)	X 等于 Y
不等于(X,Y)	X 不等于 Y
大于(X,Y)	X 大于 Y
小于等于(X,Y)	X 小于等于 Y
包含(X,Y)	X 包含 Y
不包含(X,Y)	X 不包含 Y
Advanced(X,Y)	X 比 Y 先进
交集(X,Y)	X 与 Y 的交集
Part-whole(X,Y)	X 是 Y 的一部分
纲分类$(X, Y1,Y2,\cdots Yn)$	$Y1,Y2,\cdots Yn$ 是 X 纲下的分类
Specialization(X,Y)	X 是 Y 物种的一个特例

基于植物本体的非植物结构公理使用一阶逻辑来表示。“X:〈类〉”中，X 是一个类的变量，取值为类中的任意一个实例，“$X\cdot Y$”是 X 的某个槽 Y 的取值，如“X:花类，$X\cdot$传粉方式”表示的是花类中的一个实例的传粉方式。对于关系槽，由于它本身就定义了概念与概念之间的某种关系，所以也使用本体中的这些关系槽来表示公理。如：“X:植物类，Y:植物类，进化自(X,Y)”表示的是植物 X 进化自植物 Y。除此之外，同时我们也定义了一套涵数和谓词来表示公理，除了一些在其他领域的公理中也用到的通用的谓词，我们也定义了专用于植物界的谓词表示。见表 1。

2.5.3　公理规则获取方法　植物公理和规则的获取有两种方式，一是直接从植物专业知识和常识中总结，二是从已有的公理中通过逻辑推理推出新的符合植物界规律的规则。一般以前一种为主，后一种作为前一种的辅助和补充。由于植物结构公理方面，中科院的顾芳博士已经给出了四大基本公理[2]，后面的工作都是基于这四大基本公理推出，本文作者的工作也主要是着重于非植物结构公理部分，所以在这里我们就专门讨论非植物结构公理。

2.5.3.1　从植物专业知识和常识中获取　通过总结植物学专业知识，以及与植物学专家讨论，可以总结出约束本体各个部分的公理。包括类间公理、槽间公理、值域公理、属性公理、关系公理[4,9]。

类间公理：用其中一个类来约束另一个类。对于一个类，可以找出在其植物领域上的相关类和关系，并用前面定义的谓词形式来表示。

槽间公理：一个类的槽与另一个类的槽之间的联系和约束。可以对一个类的槽在其他类中找到其有关系和比较的槽，整理出他们之间的关系。

值域公理：通过植物学专业常识，对值域的取值进行上界与下界的一些限制，使其取值和单位等在合理的范围之内。

属性和关系公理：对一个类的某种属性找出其相关属性，利用关系表示这些属性之间的联系，再通过前面定义的一些函数来表示其之间的联系。

表 2 给出我们整理出来的部分非植物结构公理。

表 2　植物公理示例(部分)

公理 1	类间公理	所有 X:植物类，所有 Y:植物类，*specialization*$(Y,X)\wedge$*specialization*$(X,Y)\leftarrow\rightarrow$等于$(X,Y)$
公理 2	类间公理	所有 X:植物类，所有 Y:植物类，进化于$(X,Y)\rightarrow$*advanced-than*(X,Y)
公理 3	槽间公理	所有 X:植物类，所有 Y:植物类，所有 Z:植物类，是组成$(X,Y)\rightarrow$包含$(X\cdot$有物质，$Y\cdot$有物质$)$
公理 4	槽间公理	所有 X:植物类，所有 Y:花类，有部分$(X,Y)\rightarrow$小于等于$(Y\cdot$寿命，$X\cdot$寿命$)$
公理 5	属性公理	植物类，包含$(X\cdot$生殖方式，$X\cdot$主要生殖方式$)$
公理 6	属性公理	所有 X:门类，小于等于$(X\cdot$已知目数，$X\cdot$已知科数$)$
公理 7	关系公理	所有 X:植物类，所有 Y:植物类，进化为$(X,Y)\rightarrow$进化于(Y,X)
公理 8	关系公理	所有 X:植物类，所有 Y:植物类，所有 Z:植物类，有物质$(X,Y)\wedge$有物质$(Y,Z)\rightarrow$有物质(X,Z)

2.5.3.2　通过公理推出其他规则　由已知的公理，我们可以通过数学逻辑计算，推导出其他相关的规则。如，有公理：X:门类，小于等于$(X\cdot$已知属数，$X\cdot$已知种数$)$，可以推出如下规则成立：

所有 X:门类，小于等于$(X\cdot$已知科数，$X\cdot$已知属数$)$

所有 X:门类，小于等于$(X\cdot$已知目数，$X\cdot$已知科数$)$

所有 X:门类，小于等于$(X\cdot$已知纲数，$X\cdot$已知目数$)$

所有 X:门类，小于等于$(X\cdot$已知门数，$X\cdot$已知目数$)$

所有 X:门类，小于等于$(X\cdot$现存门数，$X\cdot$已知门数$)$

同时，根据这条公理，还可以验证出，“规则 X:门类，小于$(X\cdot$已知种数，$X\cdot$已知属数$)$”不正确。这种方法可以在主要公理已经总结出来的情况下，对这些已有的规则进行补充和验证。

2.5.4　基于公理和规则的植物知识分析　有了植物本体的公理体系，就可以根据这些公理和规则对获取的知识进行分析。植物知识分析包括知识一致性分析和植物知识推理两个部分。一致性就是分析获取的知识是不是正确的，中间有无矛盾的地方。植物知识推理就是分析知识的内在涵义，从而得出知识库中没有的并且和植物学领域相一致的新知识[11]。假设获得如下的知识：

陆地棉，种子植物，裸子植物亚门，双子叶植物纲，锦葵科

我们就可以通过植物公理和规则推出双子叶植物纲不属于裸子植物亚门，推出这条知识有错误的信息。

植物本体构造好以后，就可以基于这个本体进行有目的的知识获取。我们现在以中科院计算所提供的文本库作为知识源，使用文本获取方法对知识进行获取[15,16]。文本库是从网络上获取的文本，以 txt 文本格式存储在磁盘上。

3 文本知识获取方法

基于植物本体的知识获取框架可以被分为两部分模块：文本过滤和知识概念获取(见图 6)。文本过滤依照植物本体从中科院提供的文本库中过滤出与植物有关的过滤文本，并将过滤文本转换成一个中间形式。知识概念获取则从过滤文本中抽取植物概念知识或关系。这里着重介绍怎样从文本中获取植物概念。

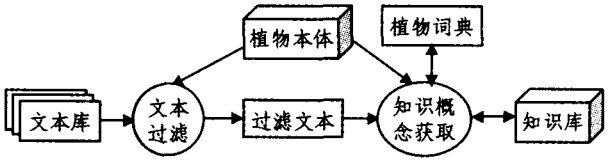

图 6 文本挖掘框架

3.1 初步过滤文本

由于文本库中的文本是直接从网络中获取过来的，是杂乱无章的，所以需要进行初步的过滤，即把与植物有关的内容先过滤出来。现在用的方法是用句型获取与植物相关的句子。这里所说的句型，也就是句子的特征形式。由于植物学中描述植物各个属性时所用的句型格式大致类似，都是很专业化的句式，它们有特定的句型特征值，如："喜生于"，"多产于"，"叶对生"等等。所以可以从网络，书本等各种材料中总结描述植物属性的句型，形成上述的句型文件。然后在文本库中过滤出符合上述句型的文本。结合对句型进行一些规则上的限制等，就可以从文本库中过滤出与植物相关的大部分文本。

从图 7 中可以看到，文件最开始定义了句子的最大长度(MAX_SENT_COUNT)和文件最大长度(MAX_FILE_LENGTH)。然后用 defconstant 定义了一些需要用到的常量，一般都是一些经过归类的虚词，代词以及符号。接下来就是我们用 defpattern 定义的句型。其中〈? c1〉代表任意字符，通过基本规则和过滤规则来对这些任意字符的长度，包含的字符集等进行限定。这样，我们就可以从文本库中找出与该句型相匹配的句子。

具体来说，对本体中的每个属性都会有一条或多条句型与其对应。每一个句型都是用来描述植物的某一个属性的。对于那些有收敛值域的属性，可以直接用值域里面的值作为句型特征值(如图 7 中的句型 002)。对于值域不收敛的，我们就需要通过总结这一类型句式的特点，抽取出能概括这些句式的特征值(如图 7 中的句型 001)。

```
MAX_SENT_COUNT 500
MAX_FILE_LENGTH 100000000

defconstant 常数
{
      标点:, |。|. |? |! |, |、|?|!|; |;
      量词:名|个|篇|种|片|块|堆|群|批|章|节|环|部分|次|步|颗|套|本|条|张|幅|款|代|缕|位|卷|册|只|双|件|台|门|棵|
株|朵|根|头|尾
      英文数字:0|1|2|3|4|5|6|7|8|9
      中文数字:一|二|三|四|五|六|七|八|九|十
      代词:这|那|这些|那些|他|他们|它|它们|她|她们|其|主要|所|本品
      疑问词:什么|哪些|何
      左引号:"|〝|'|'
      右引号:"|〞|'|'
      左括号:(|(
      右括号:)|)
}
      defpattern 句型 001
      {
            用于获取:生命周期
            句型:<?c1><为|,><?c2><生草本><?c3>
            基本规则:notbeginwith(<?C2>, <产|意|长|病|母|了|下|活|一|二|三|四|五|六|七|八|九|十>)
            过滤规则:contain(<?C2>, 一年|二年|三年|四年|多年)
       }
      defpattern 句型 002
       {
            用于获取:营养方式
            句型:<?c1><自养|厌氧|异养><?c2>
            基本规则:
            过滤规则:
       }
```

图 7 句型文件结构(部分)

这样，可以将文本库中满足句型的句子都获取出来。只要句型的特征值足够精确，就可以将文本库中描述该植物该属性的句子获取出来。但是众所周知，中文的句式是多种多样的，如果把句型的句式规则定得太死的话，会造成很多句子的遗漏，从而降低查全率，所以不能一味地追求精确率。但是如果定得太泛，又会把文本库中许多不相关的但是满足句型的句子过滤出来，造成信息冗余。所以这里就有一个平衡的关系。一个好的句型必须是查准率和查全率都很高。

通过利用句型获取文本库的文本，我们就有了第一批的过滤文件，将文本库中与植物有关的句子都获取出来，为后期的知识概念获取做准备。

3.2 植物概念的获取

有了第一批过滤文本，就可以在这个基础上进行进一步的植物概念知识的挖掘。即：从句子中获取出具体的植物概念。植物概念获取可用从以下几步进行：获取文本块、求交、获取植物概念。

3.2.1 获取包含概念的文本块 首先介绍一下什么是文本块。在过滤出来的句子中，用"/"将句型中"〈〉"中的内容分隔开来了，如"人参是/ 多/ 年生/ / 草本/ 植物，在植物分类上属五加科人参属。"我们把两个"/"符号之间的内容称为文本块，如称"草本"为句型〈? C1〉〈多|! 中文数字〉〈年生〉〈? C2〉〈木本|草本〉〈? C3〉的第五个文本块。在这个过滤文件中，我们将所有句子的第五个文本块的集合称为第五文本块集。

有了文本块的概念，我们就可以通过文本块获取列表将可能出现所需概念的文本块提取出来。如对于某个句型，其关键概念一般出现在第三个和第四个文本块，我们就可以将第三个和第四个文本块提取出来单独处理。这样做的好处就是能进一步细化目标，将一些无用的信息如标点，虚词等尽可能都过滤掉，缩小目标范围。

3.2.2 对文本块求交 由上面的介绍可以看出，已经剔除了大部分的无用知识。对于一些块，如果其内容已是我们所需要的知识了，就可以直接获取它作为目标知识。当然，这是很理想的情况。一般的情况都是具体的概念包含在文本块中，还需要进一步的过滤。对于这种文本块，我们从统计方法学的角度进行处理。

首先，对于一般的植物概念肯定会在网络文本库中出现两次以上。因为它是常见的，基本的知识概念，必然会不只一次地被引用和提到。基于这个结论，可以对文本块集进行求交。即对包含植物概念的文本块进行两两求交可以获取出具体的概念知识。但是，求交同样会求出一些无用的冗余信息，除了如标点符号，虚词，尾词等，还有因为各种原因产生的知识残缺块，如“苦”(来源于“苦菜”)，“十字”(来源于“十字花科”)等。这些都是我们所不需要的，需要剔除掉。

可以用知识获取率来判断求交结果的效率。知识获取率就是对一个文本块集来说，利用求交方式获取出有用知识的比例。如假设 A 是一个文本块集，$count(A)$是该集的文本块个数，$result(A)$是对集合 A 求交产生的有效的知识的个数，α表示知识获取率，于是就有：

$$\alpha = result(A)/count(A) \quad (1)$$

从式(1)可以看出，α的大小和查全率是成正比的。所以我们希望 $\alpha(0<\alpha<1)$越大越好，理想的情况是接近 1。另外，从 $result(A)$的定义可以看出，其值实际上就是集合 A 中出现次数大于等于两次的概念的次数。所以，要使 α 大，就必须使集合 A 中重复出现的概念增加。

从理论上，以及对文本库和求交结果等文件分析得出，要使得文本块的重复概念增加，就必须增大文本块集自身的文本块的数量。由于文本块是从文本库中抽取而来，文本块越多，就越能反映概念的出现情况。于是，我们用两种方法来增加文本块的数量。

1)加大用来过滤的文本库的量。这是最直接的办法。通过加大文本库的量，可以使过滤出来的文本也变大，这样有限的集合在更大范围的文本里出现的次数就会增多。从而使得概念重复出现的几率增多，求交的效果也就越好。

2)将属于不同句型的但又可能含有相同概念的文本块集合并。可以来看下面两个句型：

句型 1：〈? C1〉〈多|! 中文数字〉〈年生〉〈? C2〉〈木本|草本〉〈? C3〉

句型 2：〈? C1〉〈喜生于|常生于〉〈? C2〉〈地区|环境|地带〉〈? C3〉

可以看出，虽然两个句型所代表的含义不一样，结构也不同，但是两个句型的〈? C1〉中出现的都是植物的名词，所以可以将通过句型 1 获取的第一个文本块与句型 2 获取的第一个文本块合并在一起求交，这样可以在不增加文本库大小的基础上增加知识出现的概率，提高求交的效率。也即：

$$\alpha 1 = \frac{result(A+B)}{count(A+B)} > \frac{result(A)}{count(A)} = \alpha 2 \quad (2)$$

这样做可以避免由于同一个句型的局限性而产生知识遗漏的情况，提高我们的查全率。同时，也可以避免因为文本库过大而产生系统运行速度很慢的情况。

通过上述方法，对文本块集求交，再对这些结果文件进行整理，剔除掉一些如虚词、标点符号等无用的冗余词汇，再通过人机结合的方式，抽取出其中精确的具体的概念知识。

知识获取方法很多，除了以上提到的几种方法，还有如引入基于串频统计的中文分词[17]，句型相似度比较[18]，基因重组，关键词匹配[11,19]，向量模型等人工智能的方法[20]。还可以借助前人已有的成果[21,22]，使知识概念的获取更加有效。

4 实验分析

基于上面提到的各种理论方法，我们在自己的系统环境下面进行了实验。我们的机器配置如下：CPU：Pentium 4 CPU 18000MHz；内存：512MB；硬盘：50G；操作系统：WINDOWS XP；编程语言：标准 C；文本库：txt 文本库 20G。

首先进行第一轮的文本过滤。我们利用句型文件在文本库中将与植物概念有关的句型先获取出来。为了便于实验和分析，可以先只对图 7 中的句型 001(屏蔽其他的句型)进行实验。若对更多条句型进行过滤，则系统会对每条句型都产生出一个过滤文件。产生的句型文件如图 8。

```
句型：<?C1><多|!中文数字><年生><?C2><木本|草本><?C3>
禾草(/多/年生/ /草本/ )层片和根茎苔草(地下芽)层片在洲
滩植被中的建群作用最大，而这些层片在其他区域少见或不见.
鱼腥草是蔬菜的别名，三白草科，/多/年生/ /草本/ .
供：芦荟是/多/年生/百合科/草本/植物，叶片厚实，含有多种
活性物质，素有“多用良药”、“天然美容师”、“家庭医生”等美称，
可食用、观赏等多种功能.
人参是/多/年生/ /草本/植物，在植物分类上属五加科人参属.
萍果是/多/年生/ /木本/植物，因此具备树龄 treeAge 性质.
菠萝属凤梨科凤梨属，为/多/年生/常绿/草本/果品植物.
为唇形科/一/年生/或多年生/草本/植物蕾香的全草.
为唇形科一年生或/多/年生/ /草本/植物蕾香的全草.
```

图 8 过滤文件结构(部分)

从图 8 可以看出，我们已经初步将与植物概念相关的句子获取出来了，用“/”将句子分成了多个文本块。这样即完成了第一次的文本过滤。

在过滤文本中，我们对其进行文本块的抽取。获取图 8 中每个句子的第一、二、三、五个文本块，经过合并，排序等处理之后得到如图 9 的文本块结果。从图中可以看出，我们已经过滤掉了大部分的无用信息。

根据第 3 节中的理论，对文本中的块进行求交。为了验证在上节提到的提高知识获取率的方法(见 3.2.2 节)，我们使用不同大小的文本库进行文本过滤，以便产生出不同大小的过滤文本以及文本块，比较在不同的情况下产生出的植物概念有什么不同。图 10 是我们的初步求交结果文件。可以看到文件中已经包含了我们需要的信息，但是同时也有很多无用的信息，需要我们进行剔除和整理。

```
黄芩 Scutellaria baicalensis Georgi 为 多年生 草本
黄秋葵又叫羊角菜，属锦葵科 一年生 草本
鹤望兰是 多年生 草本
鱼腥草是蔬菜的别名，三白草科， 多年生 草本
魔芋是天南星科魔芋属的 多年生 草本
高山蕨菜是蕨科蕨属 多年生 草本
青贝母是专指青海产的贝母，属百合科 多年生 草本
雏菊又名延命菊，是菊科 多年生 草本
附子，学名 Ac 零 nitumSineuse， 多年生 草本
附地菜学名 Trigonotis peduncularis，别名地胡椒、黄瓜香、鸡
肠草等，为紫色科 一年生 草本
问荆为 多年生 草本
郁金香， 多年生 草本
```

图 9 获取块结果文件(部分)

黄
锦葵科葵
蔬菜
菜,,
百合科,
属科
器粟科
菜
百合，百合科百合
由
人参是
参
,
，黄黄，科属

图 10　文本块求交结果(部分)

我们统计了各种情况下的文本块数,所获取的有用的交集结果(也就是正确的植物概念)的个数,以及相对应的 α 等项的平均值(见表 3,图 11)。

表 3　文本获取数据统计

文本库大小(G)	过滤文本(k)	文本块数	知识个数	α(%)
4.72	11.2	90	8	8.9
5.375	18.15	170	28	16.5
10.32	38.05	332	79	23.8
20.64	76.1	571	180	31.5

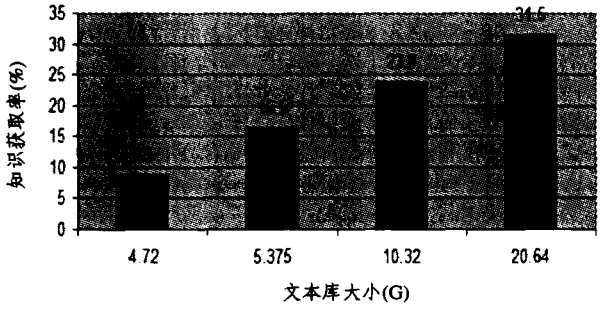

图 11　知识获取趋势图

从表 3 可以看出,随着文本库数量的变大,过滤文本也越大,获取的知识也越多,知识获取率也越高,从而证明我们的方法是正确可行的。不但如此,还可以从图 11 的知识获取趋势图上作出预测,当文本库达到 50G、100G、500G,甚至到 T 的数量级时,文本获取率会更大。可以看出,当文本库足够大的时候,我们可以获取网络文本中大部分的植物领域的知识。

通过对求交结果进行过滤,整理和排序,我们在中科院计算所提供的 20G 文本库中,仅用"〈? C1〉〈多|! 中文数字〉〈年生〉〈? C2〉〈木本|草本〉〈? C3〉"这一条句型,就获取出了 200 多条植物概念。整理后的结果如下:

人参 南瓜 倭瓜 草莓 菊花 桔梗 秋海棠 芍药 百合

图 12　植物概念获取结果文件(部分)

可以看出,这些都是所需要的植物领域的概念知识。如果将其他的句型都应用到系统,效果会更好。我们可以对这些结果进行进一步的精确化处理,验证其有效性后,再以这些概念知识为出发点,获取更多的概念。从上面的结果文件可以看出,我们的知识获取系统效率和正确率都是比较高的。

总结　植物本体是本体论在国家知识基础设施(NKI)上的一个应用,它是对植物界的特征,属性的一个全面、专业的反映和描述。我们现在已经建立了一个比较完善的植物学本体体系,并总结了一些用于校验系统的公理。基于植物学的特性和本体,我们总结了一些知识获取模式,为每个属性都总结了常见句型。目前,我们正在研究如何自动发现新的句型,以便自动扩充句型库。

由于植物学科本身的复杂性,以及在处理知识上各个人的理解的偏差,我们在知识获取,验证等方面都遇到了很多困难,这也从另一个方面对我们本体设计的合理性,完备性等方面提出了挑战。今后主要的工作将包括知识的自动获取、自动发现新句型、知识验证等方面,使我们的植物本体知识库日趋完备。

致谢　本项工作得到顾芳博士、王海涛博士生和张春霞博士生的帮助,在此表示衷心的感谢!

参 考 文 献

1　杨秋芬,陈跃新. Ontology 方法学综述. 北京大学学报(自然科学版),2002(9)

2　Gu Fang, Cao Cungen, Sui Yuefei, et al. Domain-Specific Ontology of Botany. J. Comput. Sci. Technol, 2004, 19(2): 238～248

3　The Botany Knowledge Base. Downloadable from http://www.cs.utexas.edu/users/mfkb/RKF/bkb.html

4　王丽丽,曹存根,顾芳,田雯. 基于本体论的民族知识获取与分析. 计算机科学,2003,30(5):47～51

5　曹存根. 面向专家的知识获取. 北京:科学出版社,1998

6　周云龙. 植物生物学(第二版). 北京:北京师范大学

7　中国大百科全书. 北京:中国大百科全书出版社,1998

8　陆时万主编. 植物学. 上海:高等教育出版社

9　周肖彬,曹存根. 基于本体的医学知识获取. 计算机科学,2003,30(10):35～54

10　马炜梁主编. 高等植物及多样性. 高等教育出版社

11　Maedche A, Staab S. Mining Non-Taxonomic Conceptual Relations from Text. Downloadable from http://ra.crema.unimi.it/lenya/turing/live/know/knowpubs/mining-non-taxonomic-conceptual.pdf

12　Tan Ah-Hwee, Ridge K. Digital Labs. Text Mining: The state of the art and the challenges. Downloadable from http://www.ewastrategist.com/papers/text_mining_kdad99.pdf

13　Velardi P, Fabriani P, Missikoff M. Using Text Processing Techniques to Automatically Enrich a Domain Ontology. In: Proc. of the Intl. Conf. on Formal Ontology in Information Systems, 2001. 270～284

14　Gruber T R. A translation approach to portable ontology specification. Knowledge Acquisition, 1993, 5(2): 199～220

15　Hearst M A. Text data mining: Issues, techniques, and the relationship to information access. Presentation notes for UW/MS workshop on data mining, July 1997

16　Tan A-H. Cascade ARTMAP: Integrating neural computation and symbolic knowledge processing. IEEE Trans. on Neural Networks, 1997, 8(2): 237～250

17　宋柔,许勇. 基于词汇语意的百科词典知识提取实验. Computational Linguistics and Chinese Language Processing, 2002, 7(2): 101～112

18　刘挺,吴岩,王开铸. 串频统计和词形匹配相结合的汉语自动分词系统. 中文信息学报,1998(1):17～25

19　Gomez F, Hull R, Segami S. Acquiring Knowledge From Encyclopedic Texts. In: Proc. of the Fourth ACL Conf. on Applied Natural Language Processing, Stuttgart, Germany, 1994

20　Smrz P, Rychl P. Finding Semantically Related Words in Large Corpora. TSD, 2001. 108～115

21　Marti A, Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In: Proc. of the Fourteenth Intl. Conf. on Computational Linguistic, Nantes Frances, July 1992. 9～17

22　Simoudis E. Reality check for data mining. IEEE Expert, 1996, 11(5): 26～33