

音识别软件没有中国历史知识,它怎能理解“久视一年,武曌提拔了他”中的“久视”和“武曌”的含义呢?它很可能将上述发音错误地理解为“九十一年,武照提拔了他”。由这些例子可以看出,计算机高技术产品的开发需要大量的专业知识。

不仅如此,其它高技术产品的开发也需要大量的专业知识支持。这些高技术产品包括实时网上口语翻译系统、模式识别、专家系统、大规模计算机辅助设计、计算机辅助医学、计算机辅助教学、计算机辅助决策、知识发现、分布式人工智能、自动推理、进化计算、智能计算机,等等。在建造 NKI 时,将建立 NKI 与这些系统的 API(Application Programming Interface),以便能够存取 NKI 中的知识。

5 国家知识基础设施将给传统的学校教育带来一次深刻的变革

NKI 中的知识不仅丰富,而且具有标准化和可操作的特点。这为在 NKI 的基础上大批量开发各学科和跨学科的教学设施提供了有效的支持。典型的教学设施包括网上教室、课程问题解答、网上实验室(如物理实验室和生物实验室)、学生心理咨询、可操作化的网上百科全书、网上大型博物馆、网上音乐、职业培训中心等。事实上,学生们会发现,NKI 是一个取之不尽的知识海洋。

另一方面,如果把已经发展起来的计算机动画和虚拟现实技术与 NKI 结合起来,学生们会发现,NKI 不仅是一个取之不尽的知识海洋,而且是一个活生生的知识海洋,这使学生从 NKI 中能更多地摄取知识。这里,“活生生的知识”是指知识动画和可视化(Knowledge Animation and Visualization)。很显然,这一切都是以 NKI 中的知识是可操作为必要前提的。

6 国家知识基础设施将为社会提供全方位、多样化和多层次的知识服务

上面所说的 NKI 在教学上的作用是很狭义的。从知识服务这一更广阔的角度看,NKI 可以通过相应的应用系统为社会提供全方位和多样化的服务。例如,利用 NKI 中的各种医学知识,可以建立各类医疗咨询(如婴幼儿疾病、中老年人心脑血管疾病、

癌症病人的管理)、各种法律咨询、各国地理和历史的咨询、宗教问题咨询、幼儿教育问题咨询、家庭和婚姻问题咨询以及其它专业咨询。

为能让不同文化层次的人从 NKI 中受益,需要考虑知识服务的多层次性。这里有两个有意义的问题值得研究。第一是 NKI 中的知识应有不同的抽象层次。在建设 NKI 中的知识服务器时,要将其中的知识进行必要的不同层次的抽象,以便使知识服务器能为不同用户提供不同的服务。第二是语音技术。用户可通过家用电话或移动电话与 NKI 中的知识服务器相连接,请求服务(目前国外已有此类相应的医疗咨询系统)。

我们的“梦想”就是让每个需要知识的人随时、随地、以任何方式有效地存取所需的合法知识。

7 国家知识基础设施的建设将对计算机技术的发展起巨大推动作用

NKI 是一个庞大而又复杂的知识系统,建立这样一个系统不可避免地要涉及大量的理论问题、技术问题和工程问题。以下问题值得我们从一个更高和更广的角度加以研究。

(1) 知识分类体系。各个学科的知识在数量上非常庞大,而且快速增长。因此,在问题分类的基础上建立一套完备的知识分类逻辑体系,为 NKI 中知识的分布、存取、管理提供保障。

(2) 知识的形式化。对于国家知识基础设施这样的大型知识服务项目,需要一种形式化的、多层次的、多媒体的、普适性强的知识表达体系。

(3) 知识获取。知识获取是知识工程的瓶颈问题,也是 NKI 建设的一个关键问题。我们将研究知识获取自动化的理论和技术,并重点研究从大规模文本、大规模数据库和因特网上的自动知识获取。同时,还要继续研究如何有效地获取领域专家的专业知识。

(4) 知识的模态转换。一般而言,所有的知识都是非常复杂的、维数极高的理论体系,它们难以直接满足或适用于五花八门的应用系统(如远程教学系统和自然语言处理系统)。因此,在 NKI 的总体框架下必须解决下列基本问题:过程知识到计

算机动画的自动生成; ④知识的自然语言的自动生成; ⑤知识的图形化; ⑥知识与语音合成。

(5) 矛盾知识的检测和处理。人类知识都具有一定矛盾性或可反驳性。不同的学者对某些事物的认识可能不同, 因而有些知识是相互矛盾的。研究矛盾知识处理的理论和技术, 能尽可能地发现和辨别 NKI 中的矛盾知识, 并且引导人们如何使用矛盾知识。

(6) 知识的再处理。研究知识的各种操作, 如知识抽象、求精、求同、求异、试图、组合、分解、选择、删除和添加等等, 为知识服务提供必要的基本设施。

(7) 知识的分布、传输和保护。根据 NKI 的特点, 研究分布式知识库理论和技术、多媒体知识的网络传输问题、基于因特网和 Intranet 的知识防火墙理论和技术, 等等。

(8) 开放的知识存取协议。研究整个 NKI 的构架、知识服务器的分布与互联以及为支持各类知识的应用所提供的开放的知识存取协议。同时, 还必须考虑知识存取的合法性。例如, 不是所有的人都能存取详细的人体解剖医学知识, 也不是所有的人都能存取生化武器的制作知识。

(9) 知识的更新和维护。人类对未知世界的追求是无止境的, 新的知识不断出现, 旧的理论也会不断被否定或改进。因此, NKI 中的知识需要不断更新。在建设 NKI 时, 不可避免地会将一些正确知识错误地输进 NKI 中去。因此, 知识维护也将是 NKI 生命周期中的一个重要问题。

对软件开发而言, 国家知识基础设施也将起到加速软件开发的重要作用。在社会高度信息化和知识化的今天, 各种企事业单位需要大量而又复杂的软件(如金融信息处理和管理软件)。在软件设计之初, 软件工程师必须花费大量时间来建立这样或那样抽象的软件模型(包括模型参数及其各种关系)。站在知识的层次上看, 这种抽象的软件模型代表了软件工程师对相关领域的理解, 它的绝大部分其实就是相关领域的知识模型。如果国家知识基础设施已经存在相应的知识模型, 软件工程师将节省许多时间。

8 国家知识基础设施与新一代计算机的结合

通过国家知识基础设施的建立和完善, 希望有一天将会生产出一种生物“知识芯片”。它含有国家知识基础设施中的所有知识, 而且具有自学能力。这种新一代计算机的能力将是非常巨大的。

9 国内外相关研究及分析

在建立大规模知识系统方面, 国外已开展多年的工作。下面简要介绍三项相关的工作。

(1) Cyc(1984—现在)。Cyc 是由美国的莱纳特等人正在研制的一个庞大的常识知识库系统。他们认为, 常识是研究人类智能的基础之一, 是研究人类智能本质的必要。同时为自然语言处理、语义信息检索和一致性检查等提供了共同的基础知识^[1-5]。

(2) BKB(1986—1994 年)。德克萨斯大学的伯特教授等人经过 8 年完成的植物学知识库(Botany Knowledge Base)。大约含有 18 万条事实、3 万个概念。用于自然语言处理、植物学教学等^[6]。

(3) WordNet(1985—现在)。Princeton 大学的科学家花费了大量的时间和经费研制的词汇知识系统^[7-9]。目前的 WordNet 1.6 仍有其局限。例如, 它没有 Pinta(品他病)、Treponema Carateum(品他密螺旋体)、Paris Treaty(1783 年 9 月 3 日美、英在巴黎签署的和平条约)等概念, 更没有刘少奇、周恩来、邓小平等中国领导人的名字。

另外, 还有一些比较有名的大型知识系统, 如 EcoCyc(大肠埃希氏杆菌知识系统), 到目前为止, 它收集了大量有关 E. coli 病菌方面的生物知识。

在中国, 中国科学院计算技术研究所的大规模知识处理课题组已从 12 个学科(包括中医、中国历史、世界历史、中国地理、世界地理、化工、教育学等)中获取了近 30 万个专业概念, 这是迄今为止世界上最大的多学科的专业知识库。计划再用 5 年时间, 使该知识库能覆盖所有的学科。然后, 在此基础上研究和开发高技术产品(如专业学科的机器翻译系统、语音识别系统、智能教学系统、专家系统等)。

10 结束语

人们多年来的研究成果表明,“信息”和“知识”这两个非常基本的概念之间存在着本质上的区别。我们认为,信息基础设施的主要内涵应该是层次更高的知识。如果说信息基础设施值得研究的话,知识基础设施更值得人们重视。

我们建议,将国家知识基础设施的建设列为国家发展的一项战略任务,组织我国的科研力量,全面地开展国家知识基础设施的研究和建设,以促进我国基础研究的进步、知识经济的持续发展以及科学知识的普及。

参考文献

- 1 Guha R V, Lenat D B. Cyc: a midterm report. AI magazine, 1990, 11(3): 32- 59.
- 2 Lenat D B. Cyc: A Large- Scale Investment in Knowledge Infrastructure. Communications of the ACM, 1995, 38(11): 33- 38.
- 3 Lenat D B. Steps to Sharing Knowledge. Mars N. J. I. ed. Toward Very Large Knowledge Bases. Amsterdam: IOS Press, 1995.
- 4 Guha R V, Lenat D B. Enabling agents to work together. Communications of the ACM, 1994, 37(7): 127- 142.
- 5 Lenat D B, Guha R V. Building Large Knowledge- Based Systems. MA: Addison- Wesley, 1990.
- 6 <http://www.cs.utexas.edu/users/mfkb/rkf/bkb.html>.
- 7 Lenat D B, Miller G A, Yokoi T. CYC, WordNet and EDR- critiques and responses - discussion. Communications of the ACM, 1995, 38 (11): 45- 48.
- 8 Fellbaum C. WordNet: An Electronic Lexical Database. MA: MIT Press, 1998.
- 9 Miller G, Beckwith R, Fellbaum C et al. Five papers on WordNet. 1993 Technical Report. Princeton University.

National Knowledge Infrastructure

Cao Cungen

(Institute of Computing Technology, CAS, 100080 Beijing)

Knowledge is a special product of human activity. How to share and utilize human knowledge is a significant research problem. This article discusses the concept and importance of National Knowledge Infrastructure (NKI) from the perspectives of knowledge sharing, knowledge-intensive hi-tech product manufacturing, educational technologies, etc. We also introduce related works both abroad and at in the direction of NKI.

曹存根 中国科学院计算技术研究所研究员, 博士生导师。1964 年出生。1993 年获中国科学院数学研究所博士学位, 同年进入中国科学院计算技术研究所博士后流动站。先后在澳大利亚昆士兰大学计算机系、新加坡国立大学、美国哈佛大学医学院、美国麻省理工学院计算机科学实验室、英国爱丁堡大学工作。1999 年入选中国科学院“百人计划”。目前正带领一个课题组从事国家知识基础设施的前瞻性研究。