●文庭孝 侯经川 龚蛟腾 刘晓英 汪全莉

中文文本知识元的构建及其现实意义*

摘 要 中文文本知识元的构建是解决汉语自动分词问题,实现对中文自然语言理解,并对知识内容进行操作、管理之基础。应当以汉语主题词为基础,构建中文文本知识元,建立知识元数据库,完成对知识内容的自由操作和管理。图 1。参考文献 26。

关键词 中文文本 知识元 知识元构建 知识管理

分类号 G250

ABSTRACT The construction of Chinese text knowledge elements is a basis to solve the problem of Chinese word segmentation, to realize the understanding of Chinese natural language and to operate and manage knowledge contents. We should construct Chinese text knowledge elements from Chinese subject headings, create databases of knowledge elements, and complete the free operation and management of knowledge contents. 1 fig. 26 refs.

KEY WORDS Chinese text. Knowledge element. Construction of knowledge elements. Knowledge management.

CLASS NUMBER G250

中文文本的自动切分和中文自然语言理解是目前计算机、人工智能、信息管理(信息组织与检索)、知识管理(知识组织、知识挖掘与知识发现)等学科领域面临的共同难题,也是一道难以逾越的障碍。既然无法突破中文文本的自动切分和中文自然语言理解这一瓶颈,那么是否可以通过其他方法绕过这一障碍呢?近年来中文文本"知识元"概念的提出及其构建引起了人们的极大关注。然而,通过中文文本知识元的构建能否有效解决目前汉语自动分词和自然理解所面临的难题呢?本文期望能回答这一问题。

1 中文文本知识元的提出

知识元概念的提出源于知识经济和知识管理的兴起与发展,以及信息技术的发展与普及。前者为知识元概念的提出提供了理论准备。知识经济和知识管理的兴起与发展,使知识受到人类史无前例的关注和重视,对知识本身进行科学的管理和有效的利用并实现知识增值,成为知识时代知识管理的重要目标。而要实现这一目标,首先必须完成对知识管理对象,即知识本身进行自由切分、表达、存取、组织、检索、组合等操作,所有这些知识操作与管理最终又都归结为一点,那就是必须有一个合适的恰当的知识操作与管理单元。以前以文献单元为基础进行的知识操作与管理存在很大的弊端和缺陷,就是不能直接准确反映知识内容本身,知识管理需要构建一个适合对知识内容进行操作与管理的知识单元,即知识元,这是时代的需要。后者为知识元概念的提出提供了技术条件。我们都知道,中文文本理解的核心是语词自动切分,虽然计算机技术、网络技术和人工智能技术等现代信息技术的快速发展为汉语文本的自动切分和中文自然语言理解提供了有利条件,但最终仍无法有效突破这一瓶颈和障碍。事实证明,需要将知识管理和现代信息技术有机结合起来,共同实现和完成对知识内容本身的自由操作和管理这一难题。正是沿着这样一条发展线索,人们从文献单元和信息单元逐渐发现了知识单元,即知识元,用以操作和管理知识的知识基元。

知识元应该是可以自由切分、表达、存取、组织、检索和利用知识的最小的、独立的知识单位。知识元概念的提出经历了长期的演进过程,无数有识之士为此付出了艰辛的劳动。随着现代信息技术的发展和知识管理本身的推进,国内外的许多有识之士,早就对以文献为单元的知识管理方法提出了质疑,并提出应把目标定位在知识管理上[1]。20 世纪 80 年代初,英国著名情报学家布鲁克斯提出了绘制"认知地图"的任务[2]。美国情报科学研究所研究员斯摩尔(H. Small)提出用思想"网络图"揭示重大发现,用学术思想"网络图"来表述重大发现的来龙去脉[3~4]。我国学者也围绕知识单元和知识元相继提出了"知识基因"[5]、"知识单元"[6~8]、"知识地图"[9~10]、"概念地图"[11~13]、"知识元"[14~19]、"元知识"、"知识原子"

* 本文为中国科学院科技政策与管理科学研究所和中国科学院评估研究中心博士后项目资助研究成果。

和"知识因子"[20]等概念,并进行了深入系统的研究。其中以西安电子科技大学温有奎教授为代表对中文文本知识元进行的基础性研究尤为引人注目。知识元及其构建问题一经提出即刻引起了情报学、信息管理等领域学者的高度关注和重视[21-23]。

2 中文文本知识元构建的现实意义

早在20世纪70年代后期,弗拉基米尔·斯拉麦卡来华讲学时就曾指出,知识的控制单位将从文献深化到文献中的数据、公式、事实、结论等最小的独立的"知识元"(当时他把这称为"数据元")。一旦实现知识的控制单位由文献深化到"知识元",大量文献中所包含的"知识元"及相关信息间的链接将产生极大的知识增值,从而大大推进人类对知识的利用,促进对新知识的创造,也将推动知识资源业的重大发展。

假如真正存在最小的不可分割的能够准确表达知识内容的中文文本知识构成单元,即知识元,那么目前我们必然能够有效解决如下问题:

(1)知识的自由切分与存取。如果存在构成知识的最小独立单元,那么我们应该能够脱离现有的中文文本,根据知识元对现有的知识进行自由切分,即将人类创造的知识体系分割成由知识元构成的新的知识存取体系,而不是文献知识存取体系,这样就可以以知识元为单位对知识内容体系进行自由操作。

(2)知识的自由组织与检索。如果可以以知识元为核心对知识体系进行自由切分和存取,那么当然也能够将切分出来的知识元分门别类地组织和存储起来,以便需求者能够按照知识元来检索和索取知识,并判断检索结果与所需知识是否相符。这可以说是人类历史上知识组织与知识检索的一次划时代革命。它脱离知识载体把知识组织成由知识元构成的知识元网络体系,真正实现了人们对知识内容本身进行组织与检索的长期梦想。

(3)知识的自由组合与创造。知识元构建的最重大意义可能还在于,通过知识元的自由切分与存取、自由组织与检索可以对知识本身进行自由组合与创造了。这无疑大大加速了知识增长和创造的步伐,我们现在举步维艰的知识创新就应该变得容易了。因为目前知识创造和知识创新中的许多困难都是出于无法准确定位有用知识而造成的,一旦实现了可以自由准确定位科学研究工作所需要的知识,那么一方面节省了他们的科研时间,另一方面只要将这些表达知识的知识元自由组合就能发现许多新的知识领域和研究领域,从而大大减轻科研工作者的压力。

(4)知识的准确计量与评价。知识元构建的另一个现实意义体现在,我们可以以知识元为单位对人类所拥有的全部知识进行准确计量与评价。因为到目前为止,我们仍然不知道全人类究竟拥有多少知识,又有多少知识是有用知识,有哪些知识是我们当前迫切需要的知识。用知识载体(文献或其他载体)不能准确对知识本身进行计量与评价,因为它无法避免知识的重复炮制和无用的知识混杂。如果确立了以知识元为核心的新的知识内容体系,那么一项新创造的知识是否是新知识,质量有多高,价值有多大,应该是可以精确计量与评价的。

当然,知识元构建的现实意义还远远不止这些,但是如果能够真正实现这些,那么我们对知识的自由操作和管理就已经进入了"自由王国"。文献[20]和[24]对知识元构建的现实意义做了较好的描述,认为:"知识可分解成最小的独立单元,即知识元。知识元是构成知识结构的基元。知识元的不同排列组合可构成不同知识单元,不同知识单元按照不同逻辑关系可组成不同的知识元链接。从知识元到知识单元,再到知识结构,形成不同属性的知识链。知识元之间的不同层次、不同属性、不同学科领域的链接,是实现新知识生产、知识传播及知识有效利用的核心。实现知识元自动链接是知识标引的核心。知识标引的目的是为了发现新知识。"

3 中文文本知识元构建的困难

尽管知识元的构建具有重大的现实意义,但是我们离真正实现这一目标还有很长的距离,面临的困难和挑战还十分艰巨。我们必须首先克服如下困难:

(1)知识元的不确定性。虽然我们可以从语法上对中文文本进行分割(如可以有效将中文文本分割成字、词、句、段等知识单元),但往往会影响语义知识和语用知识的理解。通过知识载体来反映知识内容本身需要有一个基本假设,即知识载体单元(如文献、词汇等)能比较真实地反映和表达知识内容本身。但事实上用知识载体表达知识内容本身存在巨大的误差,因此人们正在努力寻找能够表达知识内容本身的最小的不可分割的独立单元,即知识元。然而,由于知识是无形的,知识表达是自由的,知识自身存在的连续性和不可分割性,很难找到一个固定的"尺度"、"标准"来衡量某一知识内容的大小。从汉语及所有语言本身的特征来看,目前能找到的表达知识内容本身的最小知识单元应该属于长度不等的词汇。

(2)知识本身具有不可分割性。由于任何知识都是一个完整的有机整体,长度大小不固定,没有稳定的、最小的构成单位,也没有像物质由分子或原子构成、生物体由细胞构成的那样清晰的系统结构,所以知识本身具有不可分割性。知识的不可分割性是因为:一是知识之间没有明确的分割标识;二是没有确定的表达知识的最小知识构成单元;三是对知识强行进行分割会造成对知识的误读,破坏知识的整体结构。

(3)中文文本的不可分性。由于汉语文本书写的连续性和非标志性特征,加上汉语语词表达内在的丰富性和多样性,导致目前中文文本自动切分成为一道不可逾越的屏障。中文文本的这种不可分性,严重阻碍和制约了中文信息自动标引、中文信息计算机理解和处理、知识组织、知识检索、知识表达、知识发现、知识管理和知识计量与评价等相关领域的发展。从目前来看,对中文文本中所含知识的有效理解最终离不开对表达知识的词汇的理解,所以对汉语文本中词汇的理解是打开对中文文本自动理解的钥匙。因为在计算机人工智能技术还没有发展到可以完全模拟和理解人的思维时,知识元最终应该由数量不等的、最小单位的且具有独立表达意义的汉语词汇构成。

因此,目前我们能做的,只是对知识的载体进行自由的分割,还不能按照知识的内在逻辑和知识结构实现对知识本身进行分割、识别、理解与操作。特别是,即便我们对知识进行了分割,但我们却破坏了知识内部间存在的隐性逻辑关系和外部网络结构,毕竟对知识的理解具有很强的环境依存性。知识隐含在语法信息中,而语法信息之间的联系和各种组合是新知识的来源。因此,新知识的发现,需要有知识辨别和发现能力,即一定量的知识储备,或称为知识存量。知识元只是提供一个辨别和发现新知识的平台,并不能替代知识主体发现新知识。知识本身不可分割,通常只能被分割成为信息单元和信息元,通过不同的信息单元和信息元形式的组合来表达知识。这只是一种间接的知识表达与组合方式,离真正意义上的知识表达与组织还相距甚远。

4　中文文本知识元构建的方法

关于知识元最核心和最关键的问题还在于:知识元是什么和如何构建知识元。如果能有效解决这两个问题,那么通过知识元实现对知识内容本身进行操作与管理也就不远了。尽管温有奎教授在其著作《知识元挖掘》和有关知识元的系列论文中对知识元的概念、结构、特征、构建及标引等方面进行了系统分析与论述,并进行了初步的实证研究,得到许多专家学者的肯定。毫无疑问,其有关中文文本知识元构建的研究是开拓性的,但笔者经过深思后仍觉得有一些疑问:首先,温有奎教授在其著作和论文中多次反复提出"知识元"这一概念,认为"知识可分解成最小的独立单元,即知识元。知识元是构成知识结构的基元。"但没有明确规定,"知识元"究竟是什么。虽然有些地方提到了知识元由描述对象的特征、属性等构成,但没有说明这些特征和属性又该用什么表示,如何组合在一起呢?我们觉得其描述是不清楚的,仍然是一个非常抽象和模糊的概念,给人一种似是而非的感觉,无法准确把握。其次,笔者认为,不管知识元怎样定义,其最终还要落实到知识元的内容上。知识元的内容由什么构成呢(即知识元的内在结构如何)?知识元本身可分吗?如果知识元不可再分,那么它理应成为知识的最小构成单元。但是如果知识元仍然可分,那就说明知识元还不是构成知识内容的最小知识单元,其本身是多变的,不稳定的。当然,尽管有时需要保持知识内容本身的完整性,需要构建不同内容层次的知识元(如文献单元、知识单元、知识元、主题概念知识),但一个最稳定的、独立的、最小且不可分的知识元才具有有效表达知识所需要的基本结构和功能。例如,我们都知道分子是构成物质的最基本单元,细胞是构成生命体的最基本单元,分子、细胞的数量及其不同的组合方式(即结构)形成了丰富多彩的物质世界和生物世界,决定了不同物质和生命体之间的差异。因此,在中文文本知识元的构建上应该找到这样的"知识分子"、"知识细胞",依此为标准来衡量,目前关于知识元的研究还没有实现这一点。

综上所述,笔者认为中文知识元最合适的表达方式应该是主题树/主题概念地图知识元(因为中文文本不能自由切分,关键词不能自由提取和组合),而英文最合适的知识表达方式是关键词知识元。主题树/主题概念地图知识元表达方式能准确反映知识元之间的各种隐含的有效关联(如等级种属关系、矛盾关系、并列同一关系、簇类关系等),但主题树/主题概念地图知识元反映的是静态的知识体系,而关键词知识元反映的是动态知识体系,但不能有效反映知识元之间存在的各种隐含关联。因此,从目前的现实出发(中文文本无法实现自由有效自动切分),既然无法绕过汉语语词切分这一瓶颈,中文关键词又存在巨大缺陷,那么可以用汉语主题词来表达知识的最小构成单元,即知识元。任何一个确定的知识元最终都是由包含若干个主题词的主题概念组合而成,可以说,知识元是一个主题词集合,每个主题词都是所属知识

元集合中的一个元素，也是表达知识元的特征词。只要将某学科领域或学科主题的所有知识元用不同数量和具有不同关系的主题词完整表达出来，形成学科知识元网络体系和主题知识元网络体系，实现对知识元的自由操作和管理也是十分有效的（如图1所示）。学科知识元集、主题知识元集、类型知识元和主题词知识元构成了四个不同层次的知识元，每种类型的知识元都可以用一组主题词来完整表达在主题知识元集形成各种类型知识元，类型知识元组合形成主题知识元集合，主题知识元集合组合形成学科知识元集合。

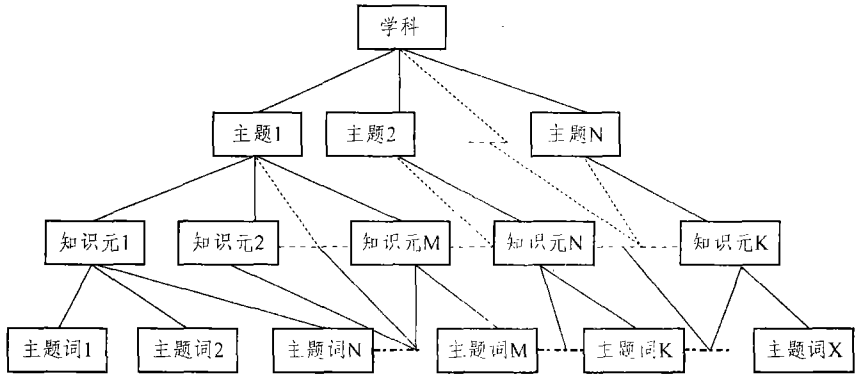

图1 知识元构建模型

当然，知识元的类型是多种多样的，表达方式也各不相同，而主题词表达知识概念的成熟性和灵活性可以完全满足知识元表达的需要。文献[20]、[24～25]将知识元分成两大类型：①描述型，包括信息报道型、名词解释型、数值型、问题描述型、文献引证型；②过程型，包括步骤型、方法型、定义型、原理型、经验型等。也可分为理论与方法知识元、事实型知识元和数值型知识元。而文献[26]将知识元分为以下几种类型：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、GDP、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、DNA双螺旋结构等。无论怎样划分知识元，但有一点必须明确，那就是知识元是构成知识网络体系的最小和最底层"节点"，这个节点本身应具备独立性、稳定性和完整性。

参考文献

1 王知津. 知识组织的目标与任务、情报理论与实践，1999(2)
2 李荫涛. 布鲁克斯认识地图初探. 情报学报，1988，7(4)
3 刘植惠. 两种新型的情报产品——《超级杂志》和《科学地图册》. 情报理论与实践，1994(6)
4,18 温有奎等. 基于知识元的文本知识标引. 情报学报，2006(3)
5 刘植惠. 知识基因理论的由来、基本内容及发展. 情报理论与实践，1998(2)
6 赵红州等. 知识单元的静智荷及其在荷空间的表示问题. 科学学与科学技术管理，1990(1)
7 王子舟，王碧滢. 知识的基本组分——文献单元和知识单元. 中国图书馆学报，2003(1)
8 徐荣生. 知识单元初论. 图书馆杂志，2001(7)
9 刘春茂. 从知识地图到数字地球——谈人类信息基础环境的演变. 中国图书馆学报，2000(5)
10 邓三鸿等. 学科知识地图的构建——以图书、情报学为例. 情报学报，2006(1)
11 马费成，郝金星. 概念地图在知识表示和知识评价中的应用(I)：概念地图的基本内涵. 中国图书馆学报，2006(3)
12 马赞成，郝金星. 概念地图在知识表示与知识评价中的应用(II)：概念地图作为知识评价的工具及其研究框架. 中国图书馆学报，2006(4)
13 马费成，郝金星. 概念地图及其结构分析在知识评价中

的应用(III):实证研究. 中国图书馆学报,2006(5)
14 温有奎. 基于知识元语义网格平台的知识发现研究. 计算机工程与应用,2006(4)
15,24 温有奎,徐国华. 知识元链接理论. 情报学报,2003(6)
16,25 温有奎,赖伯年. 网格技术将推动知识管理革命. 情报学报,2004(1)
17 温有奎等. 基于创新点的知识元挖掘. 情报学报,2005(6)
19 温有奎. 基于"知识元"的知识组织与检索. 计算机工程与应用,2005(1)
20 温有奎. 知识元挖掘. 西安:西安电子科技大学出版社,2004
21 曾民族. 向知识标引进军——阅读《知识元挖掘》的体会. 情报学报,2006(2)
22 朱庆华.《知识元挖掘》评介——兼议情报学的理论研究. 情报科学,2006(12)
23 马炳厚.《知识元挖掘》评介. 图书馆理论与实践,2006(5)
26 张静等. 论中小学多媒体知识元库的建设. 现代教育技术, 2005(5)

文庭孝 湘潭大学管理学院副教授,中国科学院科技政策与管理科学研究所博士后。通讯地址:湘潭大学管理学院。邮编411105。

侯经川 湘潭大学管理学院教授,浙江大学公共管理学院博士后。通讯地址同上。

龚蛟腾 湘潭大学管理学院讲师,武汉大学信息管理学院博士。通讯地址同上。

刘晓英 湘潭大学图书馆馆员。通讯地址同上。

汪全莉 湘潭大学管理学院讲师。通讯地址同上。

(来稿时间:2007-05-18)

(上接第64页)境的最相关的链接信息,为企业竞争情报工作提供了高质量的采集源;通过实时监控竞争对手网站,全方位地获得竞争对手最新的发展动态。还可以将Web结构挖掘方法与内容分析技术相结合,快速、多维、多层面地采集动态竞争情报。

参考文献

1 张玉峰,部先永,晏创业. 动态竞争情报及其采集基础. 中国图书馆学报,2006(12)
2 C. Aggarwal. F. Al-Garawi and P. Yu. "Intelligent Crawling on the World Wide Web with Arbitrary Predicates". In Proceedings of the 10th International WWW Conference, Hong Kong, May 2001
3 Kleinberg J. Authoritative sources in a hyperlinked environment. In: Tarjan RE, et al. , eds. Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms. New Orleans: ACM Press, 1997
4 李卫,刘建毅,何华灿等. 基于主题的智能信息采集系统的研究与实现. 计算机应用研究,2006(2)
5 Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey, SIGKDD Expirations, ACM SIGKDD, July 2000
6 朱炜,王超,李俊,潘金贵. Web超链分析算法研究. 计算机科学,2003(9)
7 李春旺. Web信息主题采集技术研究. 图书情报工作,2005(4)
8 Chakrabarti S, Dom B, Gibson D, Kumar S, Raghavan P, Rajagopalan S, Tomkins A. Experiments in topic distillation. In: Proceedings of the ACM SIGIR workshop on Hypertext Information Retrieval on the Web. Melbourne: ACM Press, 1998
9 Bharat K, Henzinger M. Improved algorithms for topic distillation in a hyperlinked environment. In: Voorhees E, et al. , eds. Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval. Melbourne: ACM Press, 1998
10 王晓宇,周傲英. 万维网的链接结构分析及其应用综述. 软件学报,2003(10)
11 Serge Abiteboul, Mihai Preda, Gregory Cobena, Adaptive online page importance computation, Proceedings of the 12th international conference on World Wide Web, Budapest, Hungary, May 20-24, 2003
12 Liwen Vaughan and Guozhu Wu. Links to commercial websites as a source of business information. Scientometrics, 2004(4)
13 杨光. 链接分析在企业竞争情报活动中的应用. 图书情报工作,2005(1)
14 吴伟. 国外竞争情报软件研究. 情报理论与实践,2004(1)
15 陈萍丽. Web挖掘及其在竞争情报系统的应用. 情报科学,2003(9)

张玉峰 武汉大学信息管理学院教授,博士生导师。通讯地址:武汉。邮编430072。

吴金红　王翠波 武汉大学信息管理学院05级情报学博士研究生。通讯地址同上。

(来稿时间:2007-02-05)