第 23 卷 第 4 期
2004 年 8 月

情 报 学 报

Vol.23, № 4
August, 2004

从信息构建看未来的知识管理

化柏林

(中国科学技术信息研究所,北京 100038)

摘要 本文从信息构建出发,进而引申到知识管理与知识推理,最终提出了知识基础工程。在设计并实现自然语言的语法开发平台时,产生了构建知识库的设想。利用语法开发平台,就可以对自然语言的语法进行开发,加上一个好的算法,就可以对大规模文本进行自动分析。对分析过的句子进行内容提取,并用面向对象方法和逻辑形式进行格式化,得到以面向对象为特征的常识知识库和以逻辑命题为特征的专家知识库,这应该是知识发现、知识管理的最高层次,也是知识工程的核心。

关键词 信息构建 自然语言处理 面向对象方法 知识库 知识管理 知识推理 知识工程 逻辑命题

Exploring the Future Knowledge Management through Information Architecture

Hua Bolin

(*Institute of Scientific and Technical Information of China*, *Beijing* 100038)

Abstract Beginning with Information Architecture, the article describes how to conduct Knowledge Management and knowledge inference, and finally brings forward knowledge basic engineering. A conception of building knowledge repository is formed when designing and implementing Grammar Development Platform based on natural language. With the use of Grammar Development Platform, grammar of natural language can be developed. Furthermore, if there is a good parser, large-scale texts can be analyzed automatically. Through extracting the content from the analyzed sentences and formatting them with object-oriented method and logical inference, common sense knowledge repository characterized with object-oriented and specialist knowledge repository characterized with logical proposition have been built, which is the highest level of Knowledge Discovery in Databases and Knowledge Management, and the core of Knowledge Engineering.

Keywords information architecture, natural language processing, object-oriented method, knowledge repository, knowledge management, knowledge inference, knowledge engineering, logical proposition.

1 信息构建的理念

信息构建(IA)主要是认知方面的问题。其研究内容包括三部分:描述信息体系各组成部分之间的关系的系统体系结构;描述生产、传递和交换信息的技术体系结构;描述信息流及其管理的运行体系结构。

信息构建强调的是一种设计理念,而不是一种理论。这种理念突出信息的有序组织与布局,强调用户对信息的方便访问与容易理解。如何把这种理念运用到信息的组织与发布、知识的表达与存储,成为诸多信息构建专家关注的焦点。

随着 IA 的设计从平面向立体和多维空间的发展,信息的组织也由线性组织、平面组织到立体组织,进而到虚拟组织,不断向纵深方向发展。IA 的跨学科和跨领域特性,进一步拓展了信息技术在信息整合和知识管理领域的应用。当然,信息构建的

收稿日期:2003 年 11 月 11 日

作者简介:**化柏林**,男,1977 年生,硕士研究生,主要研究方向为自然语言处理。

发展不仅仅是空间的变化,更重要的是人们解决问题的方式的变化。信息构建为知识管理奠定了必要的基础结构[7],从 IA 向 KA(知识构筑体系)过渡顺理成章。

2 知识工程引言

信息和知识存在于信息开发链的不同层次。事实(facts)是人类思想和社会活动的客观映射。数据(data)是事实的数字化、编码化和序列化。信息(information)是数据在信息媒介上的映射。知识(knowledge)是对信息的加工、吸收、提取、评价的结果。智慧(wisdom)是运用知识的能力。创新(innovation)是发展社会生产力的新智慧[7]。自然语言属于数据,文本文献和网页内容属于信息,知识库① 里存储的却是知识。知识工程处理对象来源是数据,处理的是信息,得到的是知识。在知识提取的过程中,运用了一定的智慧。因此知识工程下到数据、上到智慧,跨越多个层次。

知识按存在形式分为显性知识与隐性知识,按功能分为常识知识与专家知识。表示常识知识有以下困难:事实知识库需要的事实数据极其巨大;没有良好的定义使我们能够控制独立于其他部分的边界;常识世界的概念化将可能涉及到很多实体、功能的关系;关于某些主题的知识很难通过声明语句来获取;我们用来描述世界的很多语句仅仅是一个大概[1]。因此,用面向对象知识库来表达与存储常识知识是一个不错的选择。

信息构建力图从显示的角度清晰地刻画出信息的结构与关系,而知识管理应当以隐性知识显性化、无序知识有序化、泛化知识本体化为目标。如果说网站的设计是信息的有序表达,那么面向对象和逻辑命题就是知识的最佳存储。集知识管理、知识发现和知识推理于一体的知识工程,旨在建立面向对象知识库和逻辑命题知识库,以最贴近自然的方式来描述自然界的事物,以人们可认知计算机可理解的方式描述事物之间的规律,以便能够有效地解决信息泛滥、信息爆炸等问题;可以对重复的信息进行滤重、筛选,得到最能反映事物本质及自然规律的清晰有序的知识。

知识工程也在向着知识表达清晰化、数据组织有序化、内容存储本体化的方向发展,随着自然语言处理的新进展、面向对象方法的成熟应用,特别是本体论思想的引入,为知识工程的发展指明了方向,为知识工程的实施注入了新的活力。集知识发现、知识管理与知识推理于一体的知识工程,必将引领新一代的信息科学革命。

3 知识工程总体框架

知识工程是在自然语言处理平台的基础上,综合多种数据资源,运用多种方法与手段,旨在构筑一个能使知识表达清晰、存储有序、使用方便的庞大的知识体系。

自然语言是信息与知识的最恰当表达。人们在表达信息时,用得最多的也是自然语言。因此,不论是智能检索、自动分类还是机器翻译、知识工程,都需要让计算机理解人的语言。要让计算机理解人类语言,首先要建立一个自然语言处理平台,在此基础上开发出一部面向实用的语法,然后再进行文本分析,实现上层应用。

有了这样的平台和基础,就可以对大规模文本进行分析,得到文本描述的主题,抽取对象、对象的特征以及内在的逻辑与规律,分别追加到面向对象为特征的常识知识库和以逻辑命题为特征的专家知识库。还有一些自然语言现象的统计(而不是词语的统计),是在语法理论模型基础上对语法现象进行统计归纳,以得到自然语言的规律性的东西,是一个自学习的过程。运用线性逻辑进行语义消歧,提高自然语言理解的准确度。知识工程的系统结构见图 1。

在自然语言处理平台的基础上,方可进行知识工程的实施。知识工程由知识发现、知识管理与知识推理三大模块和面向对象知识库以及逻辑命题知识库两大知识库组成。知识发现(Knowledge Discovery in Databases, KDD)被 Fayyad 定义为从数据集中识别出有效的、新颖的、潜在有用的,以及最终可理解的模式的非平凡过程[3]。因为数据与知识的差别,因此现在越来越多的人认为数据挖掘只是知识发现的一个过程、一个阶段,数据挖掘与知识发现无论从处理对象、分析过程、最终结果和研究方法都有所不同。KDD 是从数据库的数据中识别有效的、新颖的、潜在有用的,以及最终可理解的模式的非平凡过程,而从大量文本中抽取事物的描述并非完全如此。从“若橘子红了,则橘子熟了”中识别出“橘子的

① 本知识库与主题无关,因此避免使用数据仓库或知识仓库等概念。

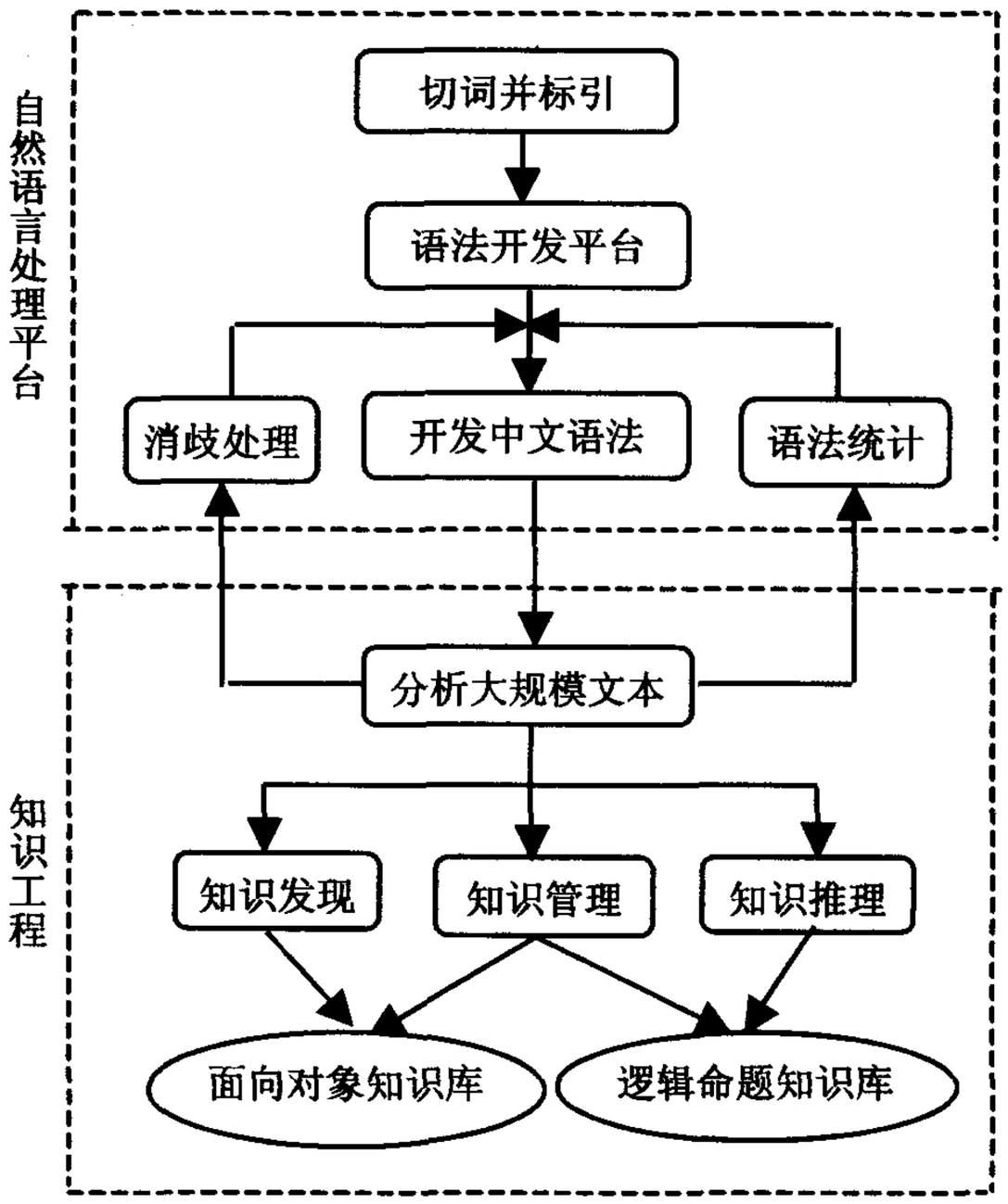

图 1　知识工程系统结构图

颜色是红色的”,这样的知识是有效的、潜在有用的、最终可理解的,但不是新颖的,是关于橘子这个对象的基本属性。因此这样一个过程叫知识萃取或知识提取会更合适。而根据“若橘子红了,则橘子熟了”这样一条规则和“橘子有点红”这样一个事实推出“橘子有点熟”这样一个结论,就是一个模糊假言的知识推理过程。

规则:若橘子红了,则橘子熟了;
事实:橘子有点红;

结论:橘子有点熟。

知识管理是创造、组织和应用知识的过程[8]。知识管理不仅是获取、组织与检索信息的问题,还涉及数据挖掘、文本聚类、数据库与文档等问题。知识与人类认知的密切相关性,决定了知识管理定位在错综复杂的结构化的内容处理上。企业级知识管理是结合某具体应用的,是专业的、面向主题的,而知识基础工程是和具体的应用无关的,是开放的、通用的。

4　建立自然语言处理平台

真正地实现知识工程,尤其是知识发现与知识推理,离不开自然语言处理平台。基于词频分布与词语共现统计的技术并不能满足对自然语言深层次分析的需要。要对文本进行理解,必须从语法语义层次上进行分析。而要能够很好地分析,就要有一部完整的语法,把语法当作一个工程来开发,需要有一个平台。语法开发平台系统结构如图 2 所示。

语法工程是一项极其巨大而且比较复杂的项目,所以只能分阶段、分领域进行,不断地加大语料的覆盖面、语种的兼容性,并从中抽出核心数据,循序渐进地生成完整的语法。该过程也是从手工到半自动,最终实现自动化的一个渐进过程。

语法开发平台利用最初的词典(lexicon)、规则(包括词规则和语法规则),对语料进行手工切分并标注,利用分析算法(parsing)生成该语种的词典、规则(包括语法规则和语义规则)。通过 LFG 的语法分析器得到成分结构和功能结构,再通过语义分析器得到句子的逻辑结构。以后在分析大规模文本时验证语法的合理性,并进一步修改、扩充与完善规则语法。开始是辅助开发语法,不断地循环验证算法,扩充与修改语法,从而得到一个完善的语法开发平台,以开发大规模实用语法。

有了语法,就可以统计大量的自然语言现象,并总结规律,完善语法模型,通过合取与析取等谓词逻辑运算进行消歧,加上一个上层应用的系统接口,这样一个自然语言处理平台就建成了。自然语言处理平台是在语法开发平台的基础上,利用完整的一部语法,加上统计模块和消歧器,外挂一个应用接口的基层平台。有了这样的平台,就可以继续构建智能检索、机器翻译、知识发现等具体应用系统。

5　用面向对象方法建立面向对象知识库

建筑出现钢结构,软件出现组件化,信息的组织出现构建。知识管理向本体化方向发展。知识工程需要从以自然语言表达的泛化文本中抽出关于事物本质的描述,包括事物的属性、构成、行为以及有关于事物的规律。

把网页或其他文本中的每一句话作为一个处理单元,借助词表对文本信息进行切词、词性标注,利用 LFG(Lexical Functional Grammar)对标注的文本进行分析,得出句子的主题与描述信息。再利用知识库中的逻辑命题对句子的 F-结构(F-Structure)进行深层次挖掘,得到更多的信息。例如通过分析“文本

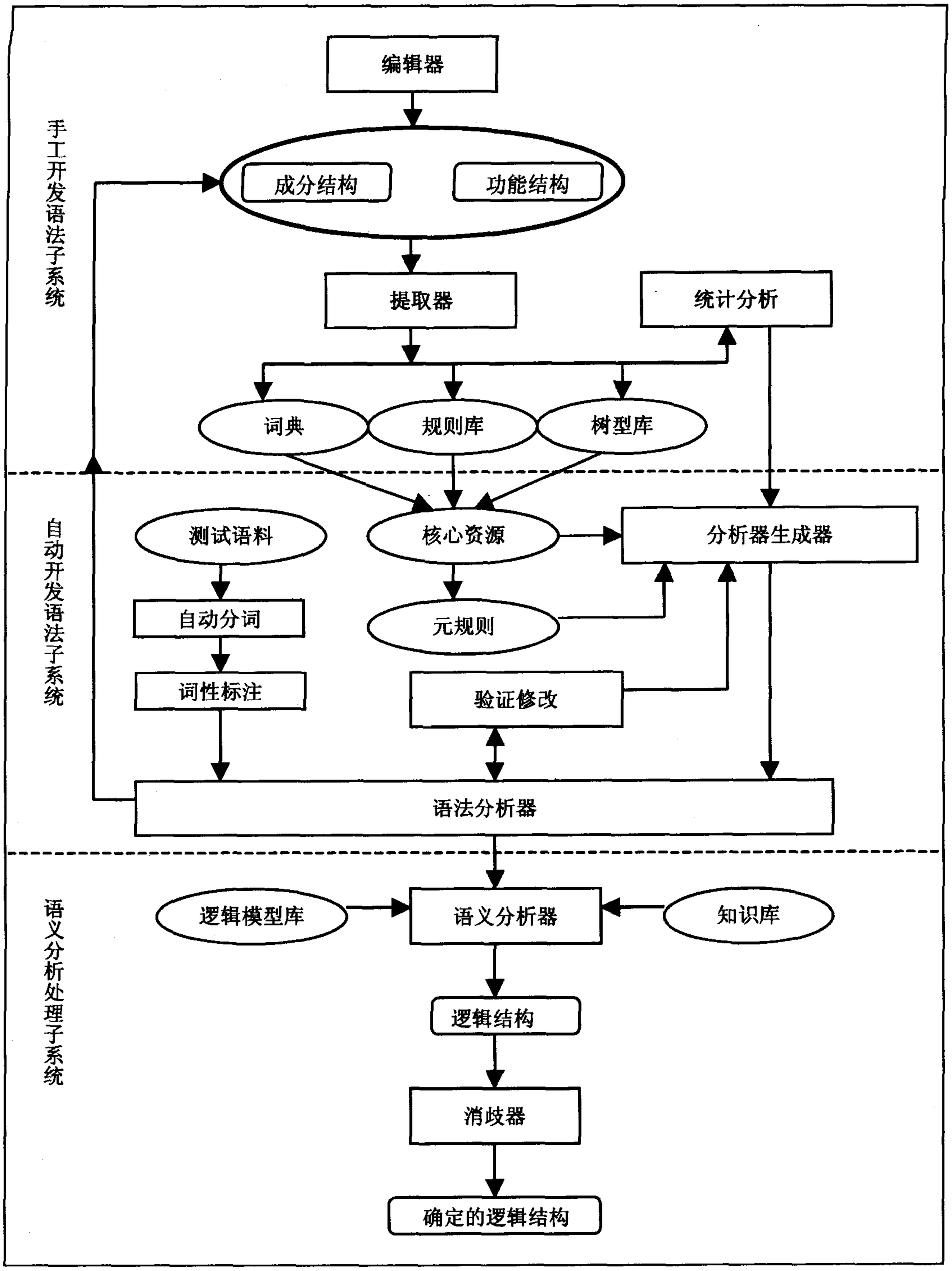

图 2 语法开发平台系统结构图[12]

文献”得出“文本是文献的一种存在形式”,而得出的结论也可以追加到以逻辑命题为特征的专家知识库中。当然,这些基于语法理论模型的研究并不能解决句间关系问题,尤其是代词所指、首语重复、词语省略和嵌入等问题[2]。要解决这些问题,至少要到语义的层次。利用语义网络的深层格结构可以分析出句子轴心的谓语与周围体词短语之间的“句法语义关系”,而用三元组集合所构建的逻辑语义网络使逻辑命题知识库的创建就更容易了。

事物的具体描述包括事物的属性及方法。分析出句子描述的对象,对象的名称、特征、功能、父子对象(上下位概念)及其关系,把这些内容封装到数据库里。用形容词来刻画属性,名词来定义构成,动词来描述方法。这样就可以从属性、构成、方法三个方

面来完整地描述一个概念。

面向对象知识库可以公式地定义对象为：O = (U,A,C,M)。其中 U 为非空的有限集论域，A 为非空的属性有限集，C 为构成对象的部件成分集，M 为对象的方法集(包括行为、功能)。

如“鸟”可以这么定义：属性有大小、颜色等；构成有翅膀、嘴巴等；行为有唱歌、飞等。在数据库里存取的是概念的抽象描述，相当于类，而分析文本时既会遇到某个具体的对象，也可能遇到类的描述。如“一头黄牛正在路边吃草”，这头黄牛是一头具体的牛，是类的实例化对象；“牛可以用来耕地”，描述的是牛的共性。

数据库的设计方案有以下几种：

(1)每一个对象作为一个记录实体，记录的值为一个属性值对 < attribute,value > 。这种存储结构有点类似 XML 文件的存储。存储空间不会浪费，根据对象名来查其属性也会很快(以对象名建立索引，用二分查找算法，速度会相当快，就像搜索引擎一样，一般为亚秒级)，但要根据属性来查相关对象，速度会相当慢。

(2)可以按照中图分类法进行分类建库，相同类目的对象具有相同的属性，存储在一个表里，再建一个关于所有表信息的表(对每个表的类进行描述)，相当于元数据表。这样建立一个标准的关系数据库，可以提高查询速度。空间浪费少，检索速度快，但归类较难。数据存储时，父类的公共属性(public 型)如果在子类中全部重新定义，造成极大的数据冗余，当然可以不定义基属性，在算法上采用继承，在存储上采用关系，但检索或使用知识时会很不方便。

无论采用何种方式，我们的目标是建立起一个以自然界的概念(包括具体的与抽象的)为对象的巨大知识库，模拟人们对万事万物的认识过程并把结果客观化、标准化、数据化。在构建面向对象知识库时，有些概念很难确定明确的隶属，可以用粗糙集来解决。

6　构建逻辑命题知识库

凡是不能用面向对象方法来存储的知识可以用命题逻辑的形式存储。知识库按内容格式分为两大类。一类以实物为特征，也就是可实例化的对象为一条记录建立数据库，记录存储着对象的属性、构成与方法；一类以逻辑命题为存储对象，每一个逻辑命题为一条记录。逻辑命题又分为多种，包括简单命题和复合命题。简单命题的真假，需要哲学、各门具体学科和人们的社会实践来确定；复合命题的真假，属于逻辑学的研究范畴，有确定的逻辑形式和逻辑关系，可以计算并有确定的答案。从建库的角度，把一阶复合命题分为联言命题、相容选言命题、不相容选言命题、假言命题、充分条件假言命题、必要条件假言命题、充分必要条件假言命题和负命题共七类，构造其真值表如下：

表 1　一阶复合命题真值表

p	q	p∧q	p∨q	要么 p，要么 q	p→q	只有 p，才 q	p←→q	~p
真	真	真	真	假	真	真	真	假
真	假	假	真	真	假	真	假	
假	真	假	真	真	真	假	假	真
假	假	假	假	假	真	真	真	

命题类型与逻辑形式是一一对应的，而命题类型与特征词却是一对多的。因此在构造命题特征词的特征关系表时，需要以命题特征词为关键字建立相应的记录。通过对一阶复合命题进行组合，就可以构建更复杂的复合命题，比如负假言命题、负相容选言命题等。

随着系统的不断完善与进展，还可以进一步构建模态逻辑库、多值逻辑库、认知逻辑库、价值逻辑库、时态逻辑库、模糊逻辑库、相干与衍推逻辑库、自由逻辑库等。逻辑命题知识库构成如图 3 所示。

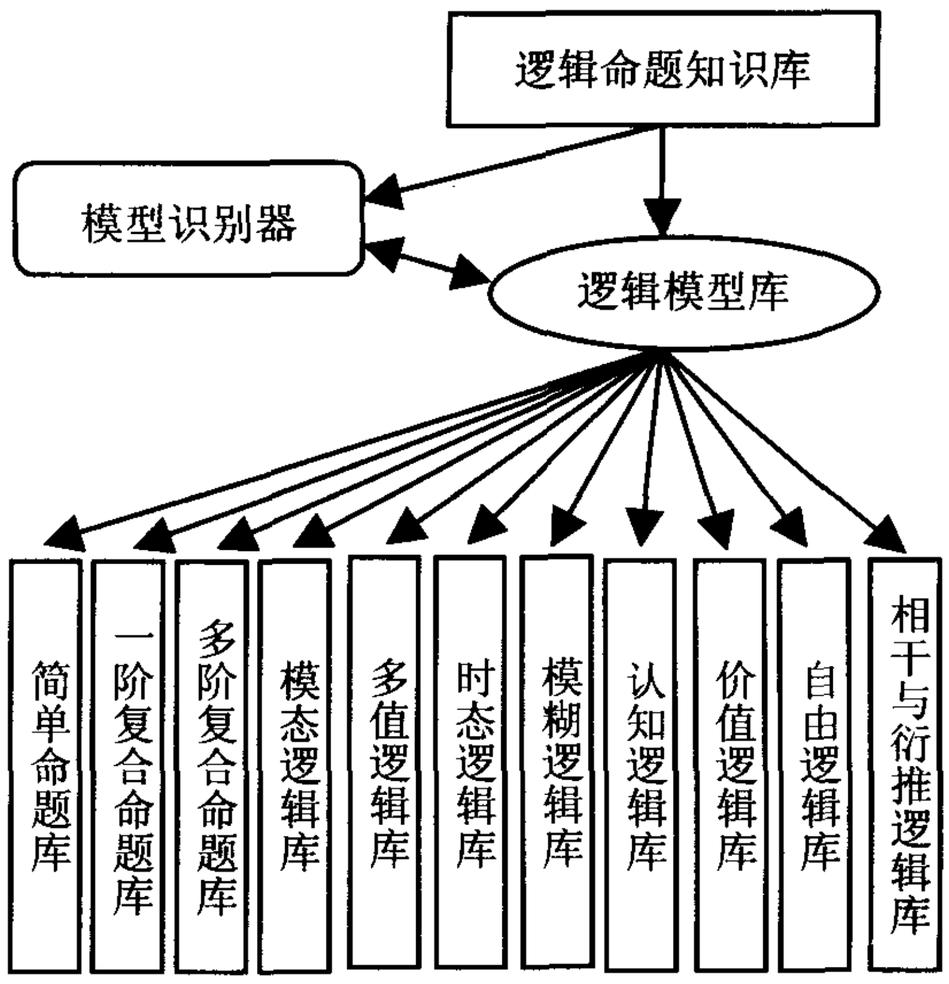

图 3　逻辑命题库构成图

逻辑库的构建可以按逻辑的对象分类建库,也可以按命题的类型分类建库。笔者更倾向后者。无论怎么分类,都要创建逻辑模型库与逻辑知识库。逻辑模型库存储的是命题类型、逻辑形式与特征词,逻辑知识库存储的是从大规模文本中抽取出的关于人们认识自然界规律的描述。通过自然语言处理,从要分析的文本中抽取出逻辑命题的特征词,到逻辑模型库进行匹配,找到相应的类型。这其实也是一个专家系统中模型的选择过程。

7 结束语

这样具有多种词典与规则库分层排布,多种知识资源库分类建立,多种方法与模型灵活选择,多个工具按序执行,多个平台协同工作的一个有机的知识工程就会构筑起来。它是一个完整的、数据异构的、具有自学习功能的、自动化、智能化的知识体系,是面向基础、体系宏伟、可实施性强的知识工程。

信息构建解决的是某个具体的信息单位(以网站为主)问题,是直接面向应用的,而知识工程的实施是一个与具体应用无关的,是通用的。信息构筑体系的形成适应了信息管理现代化、科学化和体系化的社会发展客观要求,而知识工程的实施也是一个先树后林、从具体到抽象、从局部到体系的逐步发展过程,说这样的知识库是个巨大的航空母舰一点都不过分。这样的知识工程涉及到情报学、计算语言学、逻辑学、认知学、人工智能、数据库、面向对象方法、管理学等多门学科,涉及面之广、跨学科之多、实现难度之大、开发周期之长都是情报学界罕见的,当然它的现实意义与理论高度也是无与伦比的。

参 考 文 献

1 Nils J.Nillson 著,郑扣根等译 . 人工智能 . 北京:机械工业出版社,2002

2 姚天顺 . 自然语言理解——一种让机器懂得人类语言的研究 . 北京:清华大学出版社,2002

3 史忠植 . 知识发现 . 北京:清华大学出版社,2002

4 陆汝钤 . 世纪之交的知识科学与知识工程 . 北京:清华大学出版社,2001

5 陆汝钤 . 知识科学与知识计算 . 北京:清华大学出版社,2003

6 石纯一,廖士中 . 定性推理方法 . 北京:清华大学出版社,2002

7 刘强,曾民族 . 信息构筑体系及其对推动信息服务业进步的影响 . 情报理论与实践,2003(1):1~7

8 罗威 . 在知识管理系统中实现文本挖掘 . 第十六届全国计算机信息管理学术研讨会论文集,140~146

9 苏贵阳,王永成,马颖华 . 信息基地的构架和建设模型 . 情报学报,2003(4):482~487

10 任皓,苏新宁,孔敏 . 论企业知识资源的组织 . 情报学报,2003(2):211~216

11 Lou Rosenfeld, Peter Morville. Information architecture for the World Wide Web, 2nd Edition. http://www.fucinaweb.com/design/download/iaftwww/ch09.pdf, 2003-08

12 Martin Volk Michael Jung, Dirk Richarz. GTU—A workbench for the development of natural language grammars. http://www.ifi.unizh.ch/groups/CL/CLpublications.html, 2003-08

13 Ronald M. Kaplan, Tracy Holloway King and John T. Maxwell III. Adapting existing grammars: the XLE experience. http://www2.parc.com/istl/members/thking/refs.html, 2003-05

14 J. C. Thomas, W. A. Kellogg, T.. Erickson. The knowledge management puzzle: Human and social factors in knowledge management. http://www.research.ibm.com/journal/sj/404/thomas.pdf, 2003-06

(责任编辑 许增棋)