百科知识工程*

田 野[1]，王渝丽[1]，刘 康[2]，赵 军[2]

[1]中国大百科全书出版社，北京 100037

[2]中国科学院 自动化研究所 模式识别国家重点实验室，北京 100190

E-mail: {ty, wangyuli}@ecph.com.cn; {jzhao, kliu}@nlpr.ia.ac.cn

摘 要：本文介绍百科知识工程的相关研究进展。百科知识工程的目的是在建立百科全书知识标注体系的基础上，对专家版《中国大百科全书》进行知识元标引和知识点标引等结构化处理，建立结构化的百科知识库，从而支持更为智能化的百科知识服务，同时为海量出版领域的开发利用做出示范。

关键词：百科全书；领域知识库；语义分析；知识工程

The Construction of Encyclopedia Knowledge Base

Tian Ye[1]，Wang Yuli[1]，Liu Kang[2]，Zhao Jun[2]

[1]Encyclopedia of China Publishing House, Beijing 100037

[2]National Laboratory of Pattern Recognition, Institute of Automation, The Chinese Academy of Sciences, Beijing 100190

E-mail: {ty, wangyuli}@ecph.com.cn; {jzhao, kliu}@nlpr.ia.ac.cn

Abstract: This paper gives a simple introduction to the project of Encyclopedia Knowledge Engineering. In the project, we firstly construct a knowledge representation system for Encyclopedia-XML, then we transfer the unstructured Chinese Encyclopedia into structured encyclopedia knowledge base through concept indexing and topic indexing based on the system. The knowledge base could be a very useful resource for the various tasks of intelligent information processing. The project will be a good demonstration for knowledge mining and knowledge service of tremendous publications.

Keywords: encyclopedia; knowledge base; semantic analysis; knowledge engineering

1 引言

《中国大百科全书》是我国第一部现代综合性百科全书，涉及了80多个学科，10万多个条目，有近2亿个汉字，由2万多名专家、学者、编辑经过30年的精心编纂而成，是中华文化的知识宝库。出版十余年来，受到了国内外读者的广泛关注，已经成为广大读者查询知识的必不可少的工具书。社会信息化建设的快速发展，对大百科全书的信息服务模式提出了新的挑战。用户查阅纸质大百科全书很不方便，而目前已有的大百科全书光盘版本所提供的导航式信息服务模式满足不了广大网络用户的信息需求和知识需求。如何对大百科全书进行深入的知识挖掘，并提供人性化的知识服务模式，成为百科全书发展的重要任务之一。

除了面向人的知识服务，百科全书的另外一种潜在的重要功能是面向机器的知识服务。领域知识库是智能信息处理的重要基础资源，作为领域知识库的一种重要形式，是百科知识工程的建设的重要意义。但是，百科全书目前的规模和形式远远不能满足智能信息处理应用的需求。知识工程的发展已有几十年的历史，由于传统的知识工程建设往往依赖于专家构建，其规模和更新速度不能满足真实应用的需求。网络技术的迅猛发展为知识工程的建设带来的新的机遇，以用户协作方式构建的网络百科全书以及海量的网页资源为知识工程的规模化建设提供了物质基础，而文本分析和信息抽取技术的发展也为知识工程的自动建设提供了技术手段。如何利用文本分析和信息抽取技术自动地从人类百科全书、网络百科全书和其他网络资源中抽取信息，建设可以支持计算机使用的、海量规模的、更新及时的百科知识工程成为一个重要的任务。

* 本文受国家自然科学基金项目 (60875041, 60873156)和中国出版集团科技项目资助。

以上两方面的需要涉及到以下几个关键问题：

（1）百科知识工程的知识表示：与网络百科全书相比，传统的专家编制的百科全书在精确性和系统性方面具有不可比拟的优势。如何从人类百科全书为基础建立知识描述体系，并在此基础上依据知识描述体系对网页信息进行语义标注，从而构建的开放的百科知识工程是一个重要的问题。

（2）百科知识工程的知识获取：人类百科全书、网络百科全书和其他海量网页资源是百科知识工程的重要知识来源，如何从这些多源异构的知识源中自动获取知识辅助百科专家编辑百科全书是百科知识工程建设的重要技术手段。

（3）百科知识工程的存储和检索：网络时代的百科知识工程区别于传统知识工程的重要特征是海量的规模，如何对海量的知识工程进行快速和有效的存储和检索，是另一个重要的技术问题。

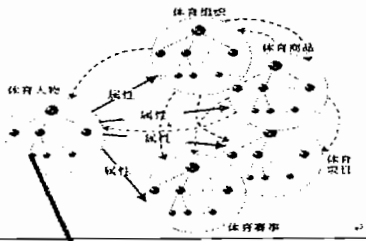
（4）百科知识工程的应用：百科知识工程具有多方面的应用，基于百科知识工程实现具有一定智能推理的百科问答系统是检验百科知识工程的一个重要应用。

为了解决以上问题，在中国出版集团的支持下，中国大百科全书出版社和中国科学院自动化研究所模式识别国家重点实验室联合进行百科知识工程建设的研究，目标是：（1）研究百科知识工程中的知识表示、知识获取、存储和检索技术，为百科知识工程的规模化、自动化建设提供核心技术支撑；（2）建设百科知识工程建设平台，支持其规模化、自动化建设，并建立规模在 300 万知识元级别的百科知识工程；（3）建立百科知识问答系统，支持百科全书的问答知识服务。

2 建立百科知识体系 Encyclopedia-XML

百科全书是权威的工具书，其编纂是个系统工程。百科的条目中包含了丰富的信息，但是目前这些信息中的大部分仅仅通过用户阅读的方式获取，缺少高性能的智能化的信息服务手段。为了能够将这些宝贵的高质量的信息以更智能化的方式提供给用户，需要能够将百科知识转化为计算机所能处理的形式。而这个转化的过程，需要有一个知识描述体系来支撑。

概念分类体系



概念实例 迈克尔·乔丹				属性描述体系
元数据	值	元数据	值	
中文名	迈克尔·乔丹	专业特点	得分、防守、领导力超强	
外文名	Michael Jordan	主要奖项	6 次 NBA 总冠军	
别名	米高·佐敦、空中飞人		2 次奥运会冠军	
国籍	美国		3 次 NBA 全明星 MVP	
出生地	纽约市布鲁克林区		5 次常规赛 MVP	
出生日期	1963 年 2 月 17 日		6 次总决赛 MVP	
毕业院校	北卡罗莱纳大学	司职	得分后卫	
身高	198cm	臂展	218cm	
体重	98.1kg	赛子	胡安尼塔	
运动项目	篮球	选秀	1984 年第一轮第 3 顺位	

图 1　百科知识描述体系示意图

《中国大百科全书》包括了 80 多个学科领域，为了充分开发利用百科知识，必须建立学科的知识体系。与网络百科全书相比，《中国大百科全书》在精确性和系统性方面具有不可比拟的优势。

为了保持百科条目的统一性和规范性，百科条目的编写者需要有一个知识体系来统一和指导它们的写作和修改。同时，百科全书的发展趋势是开放化，而在一个开放的环境下，一个高质量

的、有效的统一的规范有助于保证百科条目的高质量、完整性和描述的统一性。因此，有必要建立一个百科知识描述体系，来帮助上述两方面的处理。

整个百科知识描述体系如图 1 所示，包含知识元分类体系和属性表述体系两部分，已经由中国标准化研究院立项。在本课题中，经过学科专家和编辑人员的协作，我们首先建立了《现代医学》和《中国历史》的知识体系，进而完成了《全书》涉及的学科知识体系，为构建开放的百科知识工程奠定了基础。

3 基于 Encyclopedia-XML 的百科知识库构建平台

基于 Encyclopedia-XML 百科知识描述体系，我们建立了百科知识库构建平台，完成从百科知识库建构、百科知识库填充和百科文本自动标引等功能。

3.1 基本功能

百科知识内容加工平台的基本功能是对百科条目进行分词和词性标注。由于汉语中词与词之间没有显式的分割，因此需要有一个系统能够将其分割开来，以有助于计算机处理其内容，如索引。同时，标注词性的信息有助于用户发现文本中那些包含有实际语义信息的词，如名词和动词，同时过滤掉那些没有意义的虚词，如语气词。

给定一段百科的文本，百科知识内容加工平台首先将其切分成词，并对其进行词性标注。举例来说，给定百科中现代医学的一句话：

科学的发展在不断地加速。

百科知识内容加工平台首先按照词对其进行切分：

科学 的 发展 在 不断 地 加速 。

然后，按照北大的词性标注集对其进行词性标注：

科学/n 的/u 发展/vn 在/p 不断/d 地/u 加速/v 。/w

其中，科学被标注为名词，发展被标注为动词。中文的切分词和词性标注现在已有成熟的技术，我们也已经开发出成熟的系统，其分词和词性标注的性能都在 96%以上，能够满足实际系统的需求。

3.2 知识元标引

通常在一段文本中，包含信息最多的部分是该段文本中的命名实体。举例说来，给定如下的一段话：

1761 年，意大利人 GB.莫尔加尼出版了《疾病的位置与病因》一书。

其中的三个命名实体，*1761 年*、*GB.莫尔加尼*和《*疾病的位置与病因*》就已经提供了这段话中的绝大部分信息。

知识元标引技术，通过识别包含信息最多的部分–命名实体和术语来提取百科文本中的语义。另一方面，通过知识元的标引，我们也能够在一段百科条目的文本中提供与其内容相关的百科条目的链接，这样也有助于用户对相关信息的浏览和查找。

传统的命名实体识别技术主要识别四大类的实体：时间、人物、地点和机构。在这四个类别的实体上面，传统的命名实体识别技术已经能够达到实际应用的性能。但是，一方面百科条目还包括许多在这四个类别之外的实体，如国家、乐器、桥梁、著作和石刻等等；另一方面，仅仅将实体分为四个类别对实际应用而言还不足够，需要更进一步细化的分类，比如将人物分为帝王、后妃、神话人物、一般人物等。因此，在本系统中，我们采用传统命名实体识别技术与领域特定知识相结合的方法，来识别百科文本中的实体。

举例说来，给定一段话：

1761 年，意大利人 GB.莫尔加尼出版了《疾病的位置与病因》一书。

传统命名实体系统识别其命名实体如下：

1761 年/TIM ，/w 意大利人/n GB.莫尔加尼/PER 出版/v 了/u 《/w 疾病/n 的/u 位置/n 与/c 病因/n 》/w 一/NUM 书。/w

其中，*1761 年*被识别为时间实体，*GB.莫尔加尼*被识别为人物实体。但是，*疾病的位置与病因*未能被识别为书籍实体，因为传统的命名实体并不识别这个类别的实体。通过领域特定的识别技术，我们能够在第一步通用命名实体识别的结果上进一步识别领域特定的实体结果为：

1761 年/TIM ，/w 意大利人/n GB.莫尔加尼/PER 出版/v 了/u 《/w 疾病的位置与病因/BOOK 》/w 一/NUM 书。/w

通过标注文本中的普通命名实体和领域特定命名实体，系统一方面能够标注文本中语义信息最丰富的部分，另一方面也能够在相关的百科条目之间自动提供智能链接，如图2，在*唐太宗李世民*的条目中添加了跟其内容紧密相关的内容链接，包括*隋末农民起义*、*贞观之治*、*李靖*和*侯君集*等。

唐太宗即位后，慑于隋末农民起义的威力，居安思危，任用贤良，虚怀纳谏，实行轻徭薄赋、疏缓刑罚的政策，并且进行了一系列政治、军事改革，他的进步性政策和措施符合历史发展的规律，终于促成了社会安定、生产发展的升平景象，史称贞观之治。贞观之治是中国封建时代最著名的“治世”。　　贞观四年(630),唐太宗遣李靖平定东突厥，俘虏颉利可汗，解除了北边的威胁；九年，平定吐谷浑，俘其王慕容伏允；十四年,又派侯君集平定高昌麴氏,于其地置西州，并在交河城（今新疆吐鲁番西）置安西都护府。唐太宗对东突厥降众及依附于突厥的各族执行比较开明的政策,受到他们的拥戴,因而被尊为“天可汗”。十五年以江夏王李道宗送文成公主和亲于吐蕃的赞普松赞干布，发展了汉、藏两族间的经济文化交流。　　唐太宗是杰出的政治家，但毕竟具有阶级和历史的局限性，尤其在他的晚年，缺点暴露得越来越多。亲征高丽的战争给中朝两国人民带来了灾难，最后无功而还。　　唐太宗最初立长子承乾为太子，后来又爱重第四子魏王泰，承乾由此产生了夺嫡之惧，企图发动政变刺杀李泰，没有成功，被废为庶人。唐太宗为防止身后发生兄弟仇杀的悲剧，贬魏王泰，改立第九子晋王治为太子，即以后的唐高宗。太宗晚年著《帝范》一书以教戒太子，其中总结了他一生的政治经验，也对自己的功过进行了评述。　　贞观二十三年唐太宗病危，令长孙无忌、褚遂良在其身后辅佐李治，五月,去世，葬于昭陵。(见彩图)

图2　唐太宗李世民条目中的智能链接

3.3　元数据标引

百科专家在编写百科的每一个条目时，通常都有特定的体例来指导内容的撰写。如图 3 中宋庆龄条目的第一段，蓝色部分是关于宋庆龄的定性的语义，绿色部分是关于宋庆龄的别名的语义，红色部分是其出生信息。

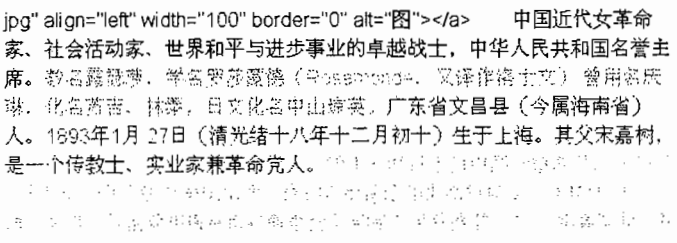

jpg" align="left" width="100" border="0" alt="图">　　中国近代女革命家、社会活动家、世界和平与进步事业的卓越战士，中华人民共和国名誉主席。[illegible]，学名[illegible]（[illegible]，又译作[illegible]）曾用名庆琳，化名[illegible]，[illegible]，日文化名中山[illegible]，广东省文昌县（今属海南省）人。1893年1月 27日（清光绪十八年十二月初十）生于上海。其父宋嘉树，是一个传教士、实业家兼革命党人。

图3　宋庆龄条目的第一段的语义

这些语义信息在许多方面都起到极为重要的作用。一方面，用户的需求通常是关于一个条目的某一个特定部分的语义，而通常基于搜索的信息服务不能够处理这部分语义信息。如果我们能够根据用户的潜在语义来查找其信息，那么准确率和召回率将有极大的提升。另一方面，对现有内容的重复出版也将大大受益于对内容的语义分析，比如为了撰写一篇关于宋庆龄的生平的文章，可以直接使用其百科条目中的生平部分，为了撰写评价宋庆龄的贡献的文章时，则百科条目中的“评价和成就”语义部分就能够重复使用。

尽管语义信息在多方面都有重要的使用，但是，目前的百科条目中并没有显式地标注这些语义信息，这使得要利用这些语义信息非常艰难。因此，我们开发出了知识点标引的关键技术，将

百科条目中的语义信息显式地标注出来：基于 Encyclopedia-XML 定义的语义规范，我们的系统能显式地标注一段文本所包含的 Encyclopedia-XML 中的语义。

人工标注语义通常能够取得高准确率，但是，由于百科知识是一个庞大的系统且具有成千上万的条目，完全人工标注将耗时耗力。为了处理这些问题，我们开发出了两套语义标注的系统：机器辅助人工语义标注系统和基于统计机器学习技术的自动语义标注系统。

自动语义标注系统需要有一定量的语料来训练模型，我们标注了中国历史卷和现代医学卷总条目的 1/3 作为训练和测试语料。目前，自动语义标注系统能够在段落级别达到 85%以上，句子级别 90%以上的准确率性能。

4 基于语义信息的问答技术

提供文档级别的大粒度搜索或是准确的句子级别或是词级别的问答，是目前智能信息服务的两种主要模式。但是，文档级别的搜索相对于百科用户的信息需求而言粒度太大：用户通常需要阅读整篇或数篇长文本才能够发现其所需要的信息，使得用户需要承担的阅读量过于庞大。另一方面，目前对于准确的问答系统的研究仅仅限制于回答一些非常简单的事实性问题，这相对于百科用户的需求而言是远远不够的。为了能够大量地减轻用户获取目标信息所需要的阅读负担，同时又能够应对用户需求的广泛性和多样性，我们开发了基于语义信息的问答技术。通过标注文本的语义和分析用户查询的语义，我们能够在语义层次发现最能够回答用户问题的答案。同时，系统能够综合在不同的文本中与用户问题相关的信息，使用户能够一次性发现所有跟问题相关的答案，即使这些答案不在一个条目内。同时，通过针对文本的不同层次进行语义标注，使得问答系统能够回答用户不同层次的问题。

针对不同文本级别的百科知识语义元数据，需要使用不同的知识标引技术。我们针对段落级和句子级的文本使用了自动分类器技术来构建百科知识自动标引系统。同时，我们进行了人机结合知识点标引，一方面，自动知识点标引的工具需要人工标注的语料来进行训练；另一方面，由于知识点自动标引的结果有时会包括少量错误，需要人工来进行校正。

软件主要包含语义标注模块，问题处理模块，答案抽取模块三个模块：

（1）语义标注模块：对百科全书的条目进行各个粒度的标注，包括条目的类别，段落及句子的语义，以及命名实体的类别标注。

（2）问题处理模块：包括题答案类型分类，问题搜索关键词处理，问题语义词检测。

（3）答案抽取模块：包括答案候选抽取，语义相似度计算，句法相似度计算。

5 工作展望

下一步的工作主要包括两方面的工作： 一方面是现有成果的推广，主要是申请百科知识描述体系 Encyclopedia-XML 国家标准；另一方面是对现有的系统进行改进和扩展。百科全书有 80 多个领域，将目前的研究成果拓展到所有这些领域，不仅仅是规模扩大的问题，如何更快速地规模化生产计算机能够处理和利用的、结构化的、语义明确的百科内容产品是一个很重要的问题。另外，在对百科全书进行知识标引的基础上开展问答式百科知识服务是我们正在进行的任务。

参 考 文 献

[1] Chien—Chung Huang, et al. Using a web-based categorization approach to generate thematic metadata from texts[J]. ACM Transactions on Asian Language Information Processing, 2004, 3(3): 190-212.

[2] Studer R, Benjamins VR, Fensel D, "Knowledge Engineering: Principles and Methods", IEEE Transactions on Data and Knowledge Engineering, 25(1-2): 161-199, 1998.

[3] A. Farquhar, R. Fikes, and J. Rice. The Ontolingua Server: A tool for collaborative ontology construction. International Journal of Human-Computer Studies, 46(6): 707-728, June 1997.

[4] H—C Yang, C—H Lee. Automatic Metadata Generation for Web Pages Using a Text Mining Approach[C]. International Workshop on Challenges in Web Information Retrieval and Integration, 2005. USA: IEEE Computer Society, 2005: 186-194.

[5] S. Dill, et a1. A case for automated largescale semantic annotation[J]. Web Semantics: Science. Services and Agents on the World Wide Web, 2003, 1(1): 115-132.

[6] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A Core of Semantic Knowledge. In 16th international World Wide Web conference (WWW2007), New York, NY, USA, 2007. ACM Press.

[7] S. Auer, C. Bizer, G Kobilarov, J. Lehmann, R. Cyganiak, and Z. G Ives. Dbpedia: A nucleus for a web of open data. In ISWC, volume 4825 of LNCS, pages 722-735. Springer, 2007.

[8] 韩先培, 赵军. 基于 Wikipedia 的语义元数据生成[J]. 中文信息学报, 2009, 23(2): 108-114.

[9] 韩先培, 齐振宇, 田野王, 渝丽, 赵军. 基于领域语义信息的百科问答系统[J]. 中国计算机语言学研究前沿进展, 2009.