国外主要知识抽取项目介绍与评析*

龚立群，孙洁丽

（中国科学院文献情报中心，北京 100080）

〔摘 要〕分别介绍国外主要知识抽取项目如 AKT、SEKT 和 ArtEquAKT 等，并进行比较和评析。

〔关键词〕知识抽取；语义网；本体；信息抽取；数字图书馆

〔中图分类号〕G250.73 〔文献标识码〕A 〔文章编号〕1002－1167（2007）04－0011－05

1 引言

目前，有大量知识存储在非结构化文档中，由于这些文档不能采用简单方式进行查询，这些文档所包含的知识既不能被自动化系统使用，也不能被人以很方便的方式进行管理，这就意味着这些知识很难得到捕获、共享和重用。知识抽取是从现有的信息（尤其是非结构化的文本）中抽取结构化的、上下文依赖的知识的过程，目的是增强信息的可使用性和可重用性[1]，这个过程同时又可以看作是对现有的非结构化信息的语义标注过程。知识抽取工具将能解析现有的文本，用其语义含义来标记文档中的概念。

语义网的初衷就是为了解决这些问题，试图给 Web 文档增加内容信息来对知识进行自动的管理。语义网目前的研究重点是如何将信息表示为计算机能够理解和处理的形式，即带有语义。因而对语义网来说，采用自动或半自动化的方法从网页上提取信息非常重要。而要实现以上目标，需要借助于本体，本体（Ontology）是共享概念模型的形式化规范说明。基于本体的语义标注工具被定义为利用已有本体在网页与文档中插入标记，或通过标注文档媒介产生知识库。利用语义标注工具对现有的大量 Web 信息进行标注，将使得 Web 页的内容成为机器可识别和“理解”的数据，从而构成语义网的基础。

数字图书馆的长远目标是从传统的信息服务转向基于知识的服务。对欧美发达国家已经完成或正在推进的一些与“知识抽取”（Knowledge Extraction）相关的研究进行调研和分析，可以为我国进行数字图书馆中知识建模和知识发现的研究提供很好的借鉴。

2 国外知识抽取项目的分类

近年来国外许多研究机构都在进行与知识抽取相关的研究项目。可以对这些项目从不同的维度进行分类。首先按与应用领域有无相关性来进行分类。有一些项目与具体的应用无关，如英国的 AKT（Advanced Knowledge Technologies）项目[2]、DELOS 的知识抽取和语义互操作（Knowledge Extraction and Semantic Interoperability）项目[3]、欧盟的 SEKT（Semantically Enabled Knowledge Technologies）项目[4]和美国的 AKDS（Automated Knowledge Discovery System）项目[5]等；有些项目是针对具体的应用领域的，如应用于医学领域的美国国家医学图书馆（NLM）的生物医学知识发现项目（Biomedical Knowledge Discovery Project）[6]和 BioMeKe[7]项目，应用于文化艺术领域的南安普顿大学的电子与计算机科学系的 Artequakt 项目[8]等。其次按项目所采用的技术实现方式来分，分为基于 Agent 的知识抽取项目，如新加坡南洋技术大学的 Intelligent Search Agent for Information Extraction and Synthesis on the Web 项目[9]和美国 KBSI（Knowledge Based System Inc）公司的 TAKE（Toolkit for Agent－based Knowledge Extraction）项目[10]；基于语义 Web Services 的知识抽取项目，如美国宇航局和阿拉巴马州立大学所进行的 SKIF（A Distributed Knowledge Extraction Framework Based on Semantic Web Services）项目[11]和欧洲的 TAO（Transitioning Applications to Ontology）项目[12]；基于遗传算法的知识抽取项目，如西班牙和阿根廷联合开展的 KEEL（Knowledge Extraction based on Evolutionary Learning）项目[13]。近年来随着多媒体资源的迅速增长，除了对传统的文本资源进行知识抽取的研究之外，国际上也出现了对于多媒体资源的知识抽取和知识标注的研究，2006 年 5 月在苏格兰的爱丁堡召开了第一届多媒体语义 Web 标注的国际会议（SWAMM06，First International Workshop on Semantic Web Annotations for Multimedia），探讨了

* 本文系国家社科基金项目“从数字信息资源中实现知识抽取的理论和方法研究”（编号：05BTQ006）研究成果之一

如何将多媒体资源的标注和语义 Web 技术结合起来[14]。因而知识抽取项目从所处理的数据源的角度来分，又可以分为基于文本的知识抽取项目和基于多媒体资源的知识抽取项目，对多媒体资源进行知识抽取的项目主要有欧盟第六框架的 K - Space（Knowledge Space of Semantic inference for automatic annotation and retrieval of multimedia content）[15]、BOEMIE（Bootstrapping Ontology Evolution with Multimedia Information Extraction）项目[16]和 CARETAKER（Content Analysis and Retrieval Technologies to Apply Knowledge Extraction to massive Recording）项目[17]等。

3 主要知识抽取项目介绍

3.1 英国的 AKT 项目

AKT 项目是一个耗资数百万英磅的由爱丁堡大学、谢菲尔德大学、南安普顿大学等大学协作开展的一个项目。AKT 的目标是，开发和提供一系列技术来解决知识工程和知识管理领域的 6 个基础瓶颈，包括知识获取、知识建模、知识重用、知识检索、知识发布和知识维护，为知识生命周期建立一套完整的方法，其中知识获取主要采用知识抽取技术，从大量的无结构的数据中，抽取出结构化的具有明确语义的知识。在项目中采用了一系列的组件技术（3Store、AKT Research Map、AKT - Bus、ANNIE、Adaptiva 等）来解决知识生命周期各个阶段所遇到的问题。目前 AKT 面临的非常具有挑战性的研究问题有：如何实现内容自动或者半自动的采集和获取；如何克服标引的瓶颈；如何了解一个用户所处的环境从而提供恰当的内容等等。

3.2 欧盟的 SEKT 项目

SEKT 项目于 2004 年 1 月 1 日开始执行，将于 2007 年结束，是一个由 EU 6th Framework（欧盟第六框架计划）资助的知识技术项目。SEKT 项目的成员来自英国电信公司等 12 个组织或大学，目标是开发和利用知识技术并以此来推动下一代的知识管理（NCKM）。下一代的知识管理系统将包括自动的知识抽取、根据用户需要进行知识打包和传递、基于语义的知识分析。SEKT 认为阻挡 NGKM 系统被广泛应用的障碍是知识的建模和知识的标注。SEKT 的三个核心技术是：Ontology 和元数据技术（OMT）、人类语言技术（Human Language Technologies）以及知识发现（Knowledge Discovery）。这三项技术将一起被应用，用于创建一系列半自动化的工具，以实现 Ontology 的创建、Ontology 中 metadata 的导入、Ontology 和相关 metadata 的演化维护。SEKT 架构将建立在国际标准之上，并且也会对新兴的语义 Web 标准产生影响。

3.3 美国国家医学图书馆（NLM）的生物医学知识发现项目

NLM 认为不论是网上提供的研究文献（特别是在 MEDLINE 数据库中），还是基因数据库中的结构化信息，它们都为生物医学领域的科学家提供了大量的、及时的、全面的信息。然而，研究者必须经过自己的详细分析和鉴别，才能知道哪些信息才是真正符合当前需求的。生物医学知识发现项目旨在研究和开发自然语言处理工具，从大量的文档中进行知识抽取，帮助科学家和其他研究人员及时跟踪和获取他们所感兴趣的主题和相关文献，并支持生物医学和分子基因领域中基于文献的知识发现。

3.4 南安普顿大学的 ArtEquAKT 项目

南安普顿大学的电子与计算机科学系的 ArtEquAKT 项目是三个项目（Artiste、Equator 和 AKT）的结合。项目的目标是使用自然语义处理技术自动地从在线文档中抽取有关艺术家的生活和工作信息，将这些信息自动地输入事先设计的领域本体中，根据用户的需求从知识库中抽取和构建信息，从而自动地产生艺术家的生平传记。ArtEquAKT 使用了自然语言处理技术来抽取关系，使用了本体来帮助处理语义信息，使用了 GATE 和 WordNet 来实现了实体识别，在术语扩展时使用了 WordNet。知识抽取工具搜索 Web 文档，并提取与给定的类目结构相匹配的知识。知识抽取工具以机器可读格式提供知识，并将这些知识存储在知识库（KB）中。该项目还未解决的问题有：（1）如何避免文档间的重复信息和重复标引。（2）知识提取虽然可以自动检索元数据模板，突破了预先定义的固定模板的限制，但却不能识别与同一类目相联系的同义词间的 Ontology 关系。（3）自动知识抽取在处理细节特征方面也存在困难。例如：可以很容易地识别出一个某一实体是指人，但识别出这个人是画家还是雕刻家却要困难一些。今后需要进一步改进的方面包括：对每个处理过程进行改进、在所产生的艺术家传记中增加图像信息、使用推理功能等。

3.5 新加坡的 Web 信息抽取和合成智能检索代理项目

Web 信息抽取和合成智能检索代理项目（Intelligent Search Agent for Information Extraction and Synthesis on the Web）是新加坡南洋技术大学的一个项目，目标是建立一个智能检索代理原型来执行 Web 上的信息抽取和合成。系统使得最终用户可以从多个 Web 站点上抽取相关信息，并将信息整合成多文档摘要。这一项目主要使用了两种技术：信息抽取、信息整合/多文档文本摘要。大多数的 Web 搜索引擎和智能检索代理只能标识 Web 上的潜在的相关文档，而不能够明确地从文档中抽取相关信息，目前所建立的信息抽取系统需要大规模的训练集，且只能够专家使用，他们的

研究寻求建立一个智能的信息抽取系统，使得普通用户使用少量的训练实例就可以使用。为了对系统的可用性进行检验，系统将应用于电子商务领域（从商务网站上自动抽取产品和价格信息）、数字图书馆、医学情报领域（从医学数据库中挖掘因果知识）。

3.6 美国的基于语义 Web 服务的分布式知识抽取框架项目

基于语义 Web 服务的分布式知识抽取框架（A Distributed Knowledge Extraction Framework Based on Semantic Web Services）是由美国宇航局和阿拉巴马州立大学联合开展的一个项目，开始于 2006 年 4 月 1 日。项目综合使用了最近几年先进的信息技术，包括：新兴的基于 Web 服务的分布式计算架构、知识工程、用来从日益增长的大量的科学观察和模型数据中实现知识抽取的科学数据挖掘等。项目的目标是建立一个语义知识整合框架原型 SKIF（semantic knowledge integration framework），SKIF 包括数据挖掘工具集、知识抽取 Web 服务和一系列相关的描述数据挖掘、管理和分析 Web 服务的本体。基于 Web 的用户界面使用本体来帮助用户发现可用的数据和服务，帮助研究者建立数据分析工作流来解决天文学领域内的问题，更重要的是可以把语义信息和知识抽取、服务的管理和分析结合起来。

3.7 欧盟的 TAO 项目

TAO（Transitioning Applications to Ontology）项目开始于 2006 年 3 月 1 日，将于 2009 年 2 月 29 日结束，总预算是 400 万欧元，来自欧洲的 5 个国家的 7 个大学或组织参加了这个项目。目标是定义一个实现将遗留系统转变为开放语义的面向服务体系架构的低成本路线，这将在异构数据源和分布应用程序间实现语义互操作。

这个项目实现了以下三方面的技术创新：（1）通过半自动化获取领域本体的语义网服务 bootstrapping，这是一个基于现有的本体学习和语义数据整合的创新性的方法。SWS bootstrapping 是目前尚未解决的一个问题；（2）扩展和整合与领域本体相关的遗留内容来实现基于本体的信息访问；（3）建立了将遗留应用程序转变为基于语义的和服务的应用程序的信息基础设施。

为了实现以上目标，项目包括一系列的研究和技术开发活动：WP1：将会形成一个涵盖 SWS bootstrapping 过程所有方面的方法。WP2：集中于从现有的应用程序文档（规格说明书、UML 图、代码文档、软件用户手册等）中学习领域本体。WP3：将会对现有遗留内容进行自动化语义扩充的方法进行研究。WP4：将会建立分布、异构知识库，以实现高效索引、查询和检索。WP5：将所有以上这些整合入一个支持转变的集成开发环境。

3.8 KEEL 项目

KEEL 是西班牙和阿根廷联合开展的一个项目，开始于 2005 年。项目的目标是为使用进化算法所建立的知识抽取模型的设计和使用提供一个计算环境。通过这个平台可以评价进化学习模型并设计新的算法。这个项目需要完成三个任务：（1）为能够集中设计和使用知识抽取模型而开发和执行 KEEL 的计算环境，建立软件库以便在将来的类似项目开发中实现快速重用；（2）分析进化学习算法现有的实验设计和检验方法；（3）建立新的知识抽取进化算法或对现有的算法进行改进。KEEL1.0 是项目所开发的一个软件工具，用来创建和使用不同的数据挖掘工具，可以说，KEEL1.0 是第一个包含有用 JAVA 所描述的所有进化学习的算法集的软件工具。KEEL1.0 包括四个组成部分：实验安装、统计分析工具、数据预处理和知识抽取算法。

3.9 Dot.Kom 项目

Dot.Kom（Designing adaptive information exTraction from text for KnOwledge Management）是由谢菲尔德大学和卡尔斯鲁厄大学等大学联合发起的一个项目[18]，开始于 2002 年 11 月。这个项目研究、设计和实施基于信息抽取的新的知识抽取方法。从科学的观点看，项目将集中于两个方面：KM 的应用给 IE 提出了怎样的需求和挑战及如何将 IE 方法转变为 KM。从实践的观点来看：项目定义了基于 IE 的 KM 工具和方法。项目来源于几个与语义网相关的领域，比较重要的有本体学习、本体维护、知识表达、文档标注和语义网服务。项目的第一个阶段已经完成，产生了几个工具：一个基于信息抽取的文档标注工具（MnM，Melita，Ontomat），一个语义浏览器（Magpie），本体管理工具（KAON 和 Ontoedit）和一个 Web 收割工具（Armadillo）。第二个阶段集中于提高和应用所建立的方法和系统。

3.10 意大利的 ONTOTEXT 项目

ONTOTEXT（From Text to Knowledge for the Semantic Web）是意大利所进行的一个项目[19]，开始于 2004 年 1 月 7 日，到 2007 年 1 月 7 日结束。目标是研究和建立创新性的知识抽取技术来产生新的、低噪音的语义网可用的信息。该项目的三个主要的研究点是：自然语言处理、基于机器学习算法的信息抽取、知识抽取和本体学习。在基于 Ontology 的知识抽取的新领域，ONTOTEXT 打算解决三个方面的问题：（1）用语义和关系信息标注文档；（2）提供足够的对这种关系信息的互操作；（3）对用来进行语义标注的本体进行更新和扩展（本体学习）。项目的预期目标描述如下：（1）在项目现有的三个理论研究领域得到提高，研究将通过一

系列的实验进行支持，研究结果将通过适合的科学渠道进行传播。(2) 大量的经过标注的文本资源将会为科学界所用。(3) 建立一个能够从报刊文章中抽取人物信息，并能够实现本体推理的系统。(4) 建立“在线人物（PEOPLE ON - LINE)”Web 服务，来提供对包含报刊文章中所提到的人物的知识库的访问。

3.11 DELOS 的知识抽取和语义互操作项目

DELOS 的知识抽取和语义互操作（Knowledge Extraction and Semantic Interoperability）项目开始于 2004 年 1 月 1 日，得到 EU 6th Framework（欧盟第六框架计划）的资助。主要目标是，针对数字图书馆中数据和描述性元数据日益增长的现状，研究并开发知识抽取和知识建模技术，完成对数字图书馆中数据的分析（对各种类型的数据如文本、音乐、统计、数学、视觉、化学、基因数据的分析）、挖掘（文本挖掘、数据挖掘、结构挖掘）、建模（建立与学位领域如经济学、数学、生物学相关的知识模型）。

3.12 欧盟的 K - Space 项目

对多媒体资源进行标注是多媒体界一直以来的一个传统，而另一方面，语义 Web 在建立基于明确的和形式化定义的语义方面已有很多经验，K - Space 项目的目标旨在将二者结合起来，从而建立起对多媒体内容的自动标注和检索的语义推理知识空间。K - Space 的创新之处在于：(1) 为低层的信号处理、对象分割、音频处理、文本分析和视听内容的构建和描述建立工具和方法；(2) 建立一个多媒体本体基础设施，分析和增强多媒体内容的知识获取、基于知识的多媒体分析、基于上下文的多媒体挖掘和用户相关反馈的智能扩展；(3) 建立多媒体内容的知识表达、多媒体数据的分布语义管理；(4) 为知识获取的协作研究建立一个开放的、可扩展的框架。

3.13 BOEMIE 项目

BOEMIE（Bootstrapping Ontology Evolution with Multimedia Information Extraction）是欧盟第六框架的一个项目，从 2006 年开始到 2008 年结束。项目的主要目标是通过引入不断发展的多媒体 ontology，为从多媒体内容中实现自动知识获取铺平道路。一方面不断地从多媒体内容中抽取语义信息来形成和丰富 ontology，另一方面，使用 ontology 又增强了知识抽取系统的功能。项目将会使用一个丰富的多媒体语义模型来产生新的抽取方法，并实现一个开放的架构。用户将能够以一种高效的方式访问知识。项目所产生的成果将会被广泛应用于商业、旅游、E - science 和新闻产业等。

4 主要项目的比较和评析

随着语义网相关技术的发展，知识抽取已成为欧美发达国家的研究热点之一。从以上介绍的项目看，国外已投入大量的人力和物力进行知识抽取项目的研究。下面对以上介绍的知识抽取项目从所属国家及地区、是否使用本体、抽取对象等方面进行比较。比较结果如下表所示。

表 主要知识抽取项目的比较

项目名称	所属国家	是否使用本体	开始时间	抽取对象
AKT	英国	使用	2001	文本
SEKT	欧盟	使用	2004	文本
生物医学知识发现项目	美国	使用	2004	文本
ArtEquAKT	英国	使用	2002	文本
Web 信息抽取和合成智能检索代理	新加坡	使用	2003	文本
基于语义 Web 服务的分布式知识抽取框架	美国	使用	2006	文本
TAO	欧盟	使用	2006	文本
KEEL	欧盟	使用	2005	文本
Dot.Kom	欧盟	使用	2002	文本
ONTOTEXT	意大利	使用	2004	文本
知识抽取和语义互操作	欧盟	使用	2004	文本
K - Space	欧盟	使用	2006	多媒体
BOEMIE	欧盟	使用	2006	多媒体

从上表可看出，所介绍的知识抽取项目都使用了本体技术，这也说明了知识抽取区别于之前传统的信息抽取，主要在于传统的信息抽取并不试图从内容上全面地、深层次地理解文档，而知识抽取则建立在信息抽取的基础之上，使用了语义网技术，从知识表示和推理的角度来实现知识的自动（半自动）抽取。目前主要的知识抽取项目大多集中于欧盟及美国，而其他国家和地区这方面的研究项目相对来说较少，亚洲较突出的研究项目是新加坡南洋技术大学所开展的 Web 信息抽取和合成智能检索代理项目。从这些知识抽取项目开始的时间来看，最早的知识抽取项目大概开始于 2002 年左右（语义网技术也正是在这一时期前后得到了快速发展），此后，欧美的许多大学和研究机构陆续开展了这方面的研究，到 2004 年、2005 年出现了许多知识抽取项目，知识抽取开始成为研究热点。而且随着基于文本的知识抽取技术及多媒体处理技术的逐渐成熟，欧美从 2006 年左右开始了基于多媒体的知识抽取研究。

5 结语

从以上对知识抽取项目的介绍和比较可以看到，随着语义网技术的逐渐成熟，知识抽取作为一门源于传统信息

抽取而又有别于信息抽取的技术，正成为信息自动化处理领域的研究热点。对欧美发达国家所推进的知识抽取项目进行介绍和分析，有助于我国进行数字图书馆中知识抽取和知识建模的研究。

〔参考文献〕

〔1〕〔2006 - 06 - 20〕. http: //delos - wp5. ukoln. ac. uk/project - outcomes/southampton/final - delos - task - 5-1-2. pdf

〔2〕〔2006 - 06 - 20〕. http: //www. aktors. org/akt/

〔3〕〔2006 - 06 - 20〕. http: //www. delos. info/index. php? option = com-content&task = view&id = 24&Itemid = 50.

〔4〕〔2006 - 06 - 20〕. http: //www. sekt - project. com/.

〔5〕〔2006 - 09 - 10〕. http: //jazz. nist. gov/atpcf/prjbriefs/prjbrief. cfm? ProjectNumber = 00 - 00 - 5509

〔6〕〔2006 - 09 - 10〕. http: //lhncbc. nlm. nih. gov/lhc/servlet/Turbine/template/branches, cgsb, KnowDiscovery. vm; jsessionid = 315B78905543B601485A78B39CB92814

〔7〕〔2006 - 06 - 20〕. http: //www. ncbi. nlm. nih. gov/entrez/query. fcgi? cmd = Retrieve&db = PubMed&list-uids = 14663967&dopt = Abstract

〔8〕〔2006 - 09 - 10〕. http: //www. artequakt. ecs. soton. ac. uk/

〔9〕〔2006 - 09 - 15〕. http: //www. ntu. edu. sg/sci/research/knowledge. html

〔10〕〔2006 - 09 - 15〕. http: //www. ntu. edu. sg/sci/research/knowledge. html

〔11〕〔2006 - 09 - 15〕. http: //aisrp. nasa. gov/projects/b8f04a0d. html

〔12〕〔2006 - 09 - 15〕. http: //cordis. europa. eu/ist/kct/tao - synopsis. htm

〔13〕〔2006 - 09 - 15〕. http: //sci2s. ugr. es/keel/description. php

〔14〕〔2006 - 09 - 16〕. http: //multimedia. semanticweb. org/

〔15〕〔2006 - 09 - 16〕. http: //cordis. europa. eu/ist/kct/kspace - synopsis. htm

〔16〕〔2006 - 09 - 16〕. http: //cordis. europa. eu/ist/kct/fp6-boemie. htm

〔17〕〔2006 - 12 - 10〕. http: //cordis. europa. eu/ist/kct/fp6-caretaker. htm

〔18〕〔2006 - 12 - 10〕. http: //nlp. shef. ac. uk/dot. kom/index. html

〔19〕〔2006 - 12 - 10〕. http: //tcc. itc. it/projects/ontotext/.

Introduction and Evaluation of Knowledge Extraction Projects Overseas

GONG Li - qun, SUN Jie - li

(Library of Chinese Academy of Sciences, Beijing 100080, China)

Abstract: Knowledge extraction is the process of extracting structured, contextually - dependant knowledge from existing information, typically unstructured text, in order to enhance the use and reuse of that information. The paper introduces overseas knowledge extraction projects, such as AKT, SEKT, ArtEquAKT, TAO, Dot. Kom, K - Space et al., and makes a comparison and evaluation of them.

Keywords: knowledge extraction; semantic Web; ontology; information extraction; digital library

〔作者简介〕龚立群（1973 - ），女，中国科学院文献情报中心 2005 级博士生。研究方向：网络信息系统建设；孙洁丽（1969 - ），女，河北经贸大学信息学院副教授，中国科学院文献情报中心 2005 级博士生。研究方向：网络信息系统建设。

〔收稿日期〕2006 - 12 - 26

（上接第 7 页）

Improvement of Community Libraries in City

ZHOU Ying - xiong

(Baoan library, Shenzhen 518101, China)

Abstract: Community libraries in city have been developing fast, playing an important part in raising the qualities of the inhabitants and accelerating the culture development. While there are problems restricting the sustaining development, we may improve the community libraries with countermeasures of law organization, model innovation, talent fostering and reasonable use.

Keywords: community library

〔作者简介〕周英雄，男，武汉大学社会学系 2006 级博士生，副研究馆员，深圳市宝安区图书馆馆长，发表论文 18 篇。

〔收稿日期〕2007 - 04 - 30