当前知识抽取的主要技术方法解析*

张智雄[1]　吴振新[1]　刘建华[1,2]　徐　健[1,2,3]　洪　娜[1,2]　赵　琦[1,2]
[1](中国科学院国家科学图书馆　北京 100190)
[2](中国科学院研究生院　北京 100049)
[3](中山大学资讯管理系　广州 510275)

【摘要】对 MnM、KIM、Text2Onto、Amilcare、Melita 等具有知识抽取功能的系统所应用的技术方法进行解析。提出在当前知识抽取技术中,机器学习和自然语言分析两大思路各自得到较大发展,并且在相互融合、相互借鉴中受益。在基于机器学习的知识抽取方面,出现以自适应信息抽取(Adaptive IE)、开放信息抽取(Open IE)为代表的新思路,并且有向自动本体学习(Ontology Learning)方向发展的趋势;在基于自然语言分析的知识抽取方面,基于模式标注、语义标注的方法得到广泛关注和进一步完善,并且有向基于 Ontology 的信息抽取(OBIE)方向发展的趋势。此外,为减少 Ontology 建设成本,让人们可以利用简单的自然语言构建 Ontology,基于受控语言的信息抽取(CLIE)技术也得到一定的关注。

【关键词】知识抽取　机器学习　自然语言分析　本体
【分类号】G250.73

Analysis of State – of – the – Art Knowledge Extraction Technologies

Zhang Zhixiong[1]　Wu Zhenxin[1]　Liu Jianhua[1,2]　Xu Jian[1,2,3]　Hong Na[1,2]　Zhao Qi[1,2]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](Graduate University of the Chinese Academy of Sciences, Beijing 100049, China)
[3](Department of Information Management, Sun Yat – Sen University, Guangzhou 510275, China)

【Abstract】Based on the analysis of some state – of – the – art knowledge extraction systems, i. e., MnM, KIM, Text2Onto, Amilcare and Melita, it brings forward that two kinds of technologies, i. e., machine learning and natural language analysis, are developed respectively and get benefits from the inter – reference. On machine learning aspect, some new methods, such as Adaptive Information Extraction, Open Information Extraction, are put forward and have a trend toward Ontology Learning. On nature language analysis aspect, the methods of Pattern – Based Annotation and Semantic Annotation get more attention than ever, and have a trend toward Ontology Based Information Extraction. Besides, Controlled Language Information Extraction method is introduced to reduce the cost of Ontology Construction and allow non – specialists to create or edit ontological data using simple nature language.

【Keywords】Knowledge extraction　Machine learning　Nature language analysis　Ontology

通常而言,知识抽取是指从数字资源中识别、发现和提取出概念、类型、事实及其相关关系、约束规则,以及进行问题求解的步骤、规则的过程。

依据数字资源类型的不同,知识抽取的概念有广义和狭义之分。广义的知识抽取泛指从各种类型的数据和信息资源中获取各种知识的过程,例如从数字信号中[1]、从多种媒体资源(如图像、数据、视频、音频)中抽取出知

收稿日期:2008 – 06 – 16
*本文系国家社会科学基金项目"从数字信息资源中实现知识抽取的理论和方法研究"(项目编号:05BTQ006)的研究成果之一。

识[2,3],从数据集中发现重要模式的过程[4]等。狭义的知识抽取则是指从非结构化的自由文本中获取相关知识内容的过程。与广义知识抽取针对各种类型数据的情况不同,狭义的知识抽取基本上属于文本挖掘的范畴,其处理的对象是自由文本,目标是分析文本内容,通过识别出文本中的知识片段(Knowledge Fragments),促进对文本内容的理解[5,6]。除纯文本文件外,狭义的知识抽取还包括对邮件、科技文献、新闻、HTML 页面、Weblogs、Wikis 等类型的数据中知识的抽取。

目前,图书馆用户所使用的数字信息资源更多是以非结构化自由文本形式存在的。狭义的知识抽取就是将这些非结构化的自由文本转换为结构化知识,以便于进一步分析和应用这些文本中的知识。狭义知识抽取是广义知识抽取的基础,也是本文关注的重点,下文中的知识抽取特指狭义的知识抽取。

知识抽取是当前自然语言处理、语义 Web、机器学习、知识工程、知识发现、文本挖掘等相关领域共同关注的重点研究之一。国内外有很多研究活动都与知识抽取相关,如英国 AKT[7]、CLEF[8]项目,欧洲 SEKT[9]、Dot. Kom[10]、DELOS[11]、X - Media[2]、OpenKnowledge[12]、K - Space[3]等项目,美国的 KnowItAll[13]、Halo[14]、RKF[15]、KXDC[16]等项目,纷纷开展从数字资源中实现知识抽取的技术研究和方法实践,研发出诸如 MnM[17]、KIM[18]、ArtequAKT[19]、Text2Onto[20]、Magpie[21]、Amilcare[22]等具有知识抽取功能的系统。

通过对当前主要知识抽取系统的分析,笔者发现当前的知识抽取系统中机器学习和自然语言分析两大技术思路正在相互融合、相互借鉴,各自都得到了较大的发展。基于机器学习的知识抽取系统,提出了自适应的信息抽取(Adaptive IE)、开放信息抽取(Open IE)等新的技术思路,并向着自动本体学习(Ontology Learning)的方向发展;而基于自然语言分析的知识抽取系统,则提出了基于模式标注(Pattern - Based Annotation)、语义标注(Semantic Annotation)等新的技术思路,并且都在向着基于 Ontology 的信息抽取(OBIE)的方向发展。另外,为了减少 Ontology 的建设成本,让人们可以利用简单的受控自然语言来构建 Ontology,基于受控语言进行信息抽取(CLIE)的技术方法也得到了一定的关注。以下将对知识抽取技术方法进行解析。

1　自适应的信息抽取(Adaptive IE)方法

Fabio Ciravegn 认为信息抽取之所以不能广泛、商业化地应用的重要因素之一是传统的信息抽取系统缺乏广泛适用性,不能实现在不同领域应用之间的快速转换[23]。为此,他提出构建自适应的信息抽取系统的设想,并开发出自适应的信息抽取系统 Amilcare[22]。Amilcare 利用$(LP)^2$规则归纳算法,借助一定数量的手工标注语料,能迅速学习标注的相关规则,以适应新的应用领域。

$(LP)^2$规则归纳算法根据用户对训练语料中不同信息内容加标记的过程总结归纳标引规则。它包括两种类型规则的归纳:

(1)标记规则(Tagging Rules)。$(LP)^2$借助训练语料中的标记实例进行语法分析和计算,生成标记规则。在这一过程中,算法根据人工标记实例和文中这一标记所在的语句,构造条件和标记关系。对于文中被用户加标记的每一句话,它以语句中的每一个词、以及这个词的语言分析结果(例如:该词的词性、大小写、是否为某个辞典中的条目、是否已经明确具有某一语义类别的词等)为条件,以标记内容为结果,自动构造出多条标记规则,并在所有语料生成的规则中,对这些规则的正确率进行计算,选择其中效果最好的 K 条规则加入到最佳规则池中,形成标记规则。为提高查全率,$(LP)^2$标记规则吸收了一些没有被纳入到最佳规则中、被称为上下文规则的规则,但上下文规则的应用有一定条件约束。

(2)修正规则(Correction Rules)。在系统利用自动学习的标记规则自动标记文本后,用户会实施人工干预,修改不正确的标记。修正规则是系统学习人们如何修改错误标记和不精确标记后形成的规则,它可以进一步提高自动标记的正确率和精确性。

Fabio Ciravegna 等在 Amilcare 基础上,开发出半自动化的标注工具 Melita[24]。利用 Melita 标注文本时,需要首先定义一个标记集(如以 Ontology 来组织),并提供需要标记的语料。用户对语料中相关文本内容加标记的同时,Amilcare 在后台运行,学习用户如何对文本进行标记,学习到的经过归纳的规则将自动应用于新文本标记过程中,利用这些规则标记出的结果可以与用户手工标记的结果对比。当这些规则的准确率达

到一定阈值后(用户可以自行定义该阈值),Melita 会自动利用规则对新文本进行预标记。此时,用户只需修改错误标记和追加遗漏标记。当然,用户修改和追加的同时,Amilcare 将继续学习修正规则。当信息抽取系统的标记输出比较可信时,用户就可以利用这一系统自动对内容进行标记了[25]。

除 Melita 外,英国 Open University 的 MnM[17] 和德国 University of Karlsruhe 的 Ontomat annotizer[26] 等很多系统也利用 Amilcare 实现了自适应的信息抽取。

2　开放信息抽取(Open IE)方法

Open IE(OIE)是美国华盛顿大学(University of Washington)图灵中心(Turing Center)提出的被称为"新型抽取范式"(A Novel Extraction Paradigm)的一种知识抽取方法[27]。Open IE 的目标在于促进领域无关的知识抽取应用,它能从文本中抽取出大量关系对,并可被应用到各种类型和规模的 Web 信息抽取任务中。除需要标注的文档集外,OIE 不需要任何其它人工输入,同时为保障在处理大规模文档集时的效率,OIE 只需要对文档集进行一次处理。

图灵中心基于 OIE 的思路,构建了名为 TEXTRUNNER[28] 的开放式信息抽取系统。TEXTRUNNER 包括三个关键模块:自监督学习器(Self - supervised Learner)、一次性通过抽取器(Single - pass Extractor)和基于冗余的评价器(Redundancy - based Assessor)[29]。

自监督学习器通过对小规模样本文献的语言分析,构造供抽取器应用的分类器。该学习器按以下两个步骤工作:

(1)通过一个完整的自然语言分析器[30]提取样本中出现的三元组 $t = (e_i, r_{i,j}, e_j)$,按一定规则将三元组标记为正值或负值。自然语言分析器对样本文献集中所有语句都进行完整的语法分析,形成语法树,找出每个句子中所有的名词短语 e_i,通过语法树构建句子中所有的名词短语对(e_i, e_j)及 $i < j$ 间可能存在的相关关系 $r_{i,j}$,从而形成一个三元组 $t = (e_i, r_{i,j}, e_j)$。对每个三元组,学习器根据这两个名词短语在语法树中是否满足某些强制性条件,将其标记为正值或负值。例如,对于一个三元组,满足以下条件:e_i 和 e_j 之间存在依赖链,且该链长度不超过某个值;在语法树中,从 e_i 到 e_j 并没有跨越句子界限(例如 e_i 和 e_j 并不是一个在主句中出现,而另一个在从句中出现);e_i 和 e_j 都不是代名词;则这个三元组被标记为正值,反之则为负值。

(2)在所有三元组都被标记后,学习器将这些三元组转换为特征向量表示,作为 Naive Bayes 分类器的输入,对 Naive Bayes 分类器进行训练。通过计算每一个特征向量正确或错误的频次,最终生成可以被抽取器应用的分类器。

一次性通过抽取器以三个步骤实现对需要标注的文档集的处理:

(1)利用轻量级的 OpenNLP Toolkit[31] 对待标注文档中每条语句进行简单的语法分析,标记出每个词的词性,并识别出名词短语;

(2)对每对名词短语,如果它们相距不远并且满足其它一些条件,则被标记为候选抽取的三元组;

(3)利用上述自监督学习器构造的分类器,对候选抽取的三元组进行分类,如果分类器认为抽取的三元组是可信的,则三元组被抽取出来,存储并归并抽取出来的三元组。最终的抽取结果中只存储各个不同的三元组和这些三元组出现的频次。为提高抽取效率,TEXTRUNNER 还及时对抽取出的结果建立索引。

基于冗余的评价器利用概率模型计算抽取出的三元组出现的频次。每一个三元组的概率可以继续被应用以提高三元组抽取的精确性。

TEXTRUNNER 目前已经对 9 000 000 个 Web 页面进行了抽取试验,得到了 11 000 000 个高概率的三元组。经过分析,这些三元组中包括 1 000 000 多个具体事实和 6 500 000 多个断言。

3　本体学习(Ontology Learning,OL)方法

基于本体进行推理获取新知识已被众多研究者证实是一种有效的知识获取方法[32],但早期本体构建工具基本上都需要人工输入大量知识,这种费时费力的任务引发了利用知识抽取技术降低本体构建开销的相关研究,即 OL。研究者们认为 OL 就是自动或半自动地从各类数据资源中获取期望本体的方法和技术集合[33],类似概念还有本体生成、本体挖掘、本体抽取等[32]。

近年来,OL 方法研究取得很大进展,但由于缺乏对 OL 具体任务的一致认定,各类方法的优劣难以比较。因此,在早期研究基础上,Philipp 等人提出将 OL 划分为专有名词、同义词、概念、概念层级、关系、关系

层级、公理模式、通用公理一系列自下而上的学习子任务(见图 1)[33]。

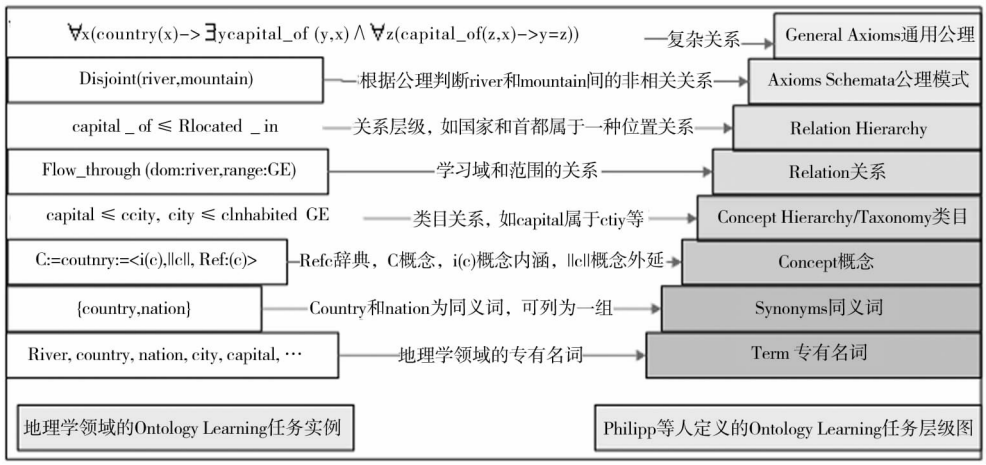

图 1　本体学习子任务层次[32]

专有名词作为领域特定概念的语言实现,识别中常采用语言学中的模式抽取、浅语义分析等方法,统计方法中的共现、频率等方法,或两者的混合方法。不同语境或语种中,专有名词往往存在同义词,通过分类、聚类等方法识别出这些同义词,可为扩展辞典提供支持。每个概念作为 < 定义,实例,同义词 > 三元组,常借助 WordNet 等辞典和形式概念分析方法实现抽取。针对多个概念间的上下位类、同位类等类目关系,研究者们较多地讨论使用辞典 - 语法模式、层级概念聚类、文档包含等方法。为构建推理规则,还需进一步识别概念间其它相关关系及各关系之间的层级,如属性关系(X of Y)、限定关系(X is used for Y)、因果关系(X leads to Y)等[34,35],较常见的方法有层级概念聚类、语义解释和关联规则等。而针对公理,目前提出的方法主要为基于模板的抽取方法。

在理论研究基础上,研究者们纷纷开发出相应工具,较为典型的有 TextToOnto 和 Text2Onto。

TextToOnto 以 KAON[36] 作为底层仓储,采用加权词频统计、概念层级聚类、关联规则和模板等方法,从非结构化数据(纯文本)和半结构化数据(HTML,词典)中获取概念及其关系,基于初始核心本体构建领域本体。

Text2Onto[37] 改进了前者依赖本体模型、缺少用户交互、缺乏动态学习等缺陷,从元数据层出发,在基于概率的本体模型中用实例模型原语的形式表达学习到的知识,整合数据驱动的变更发现策略,提高本体构建工具与本体模型的相互独立性,增强用户交互功能,实现当数据发生变化时,只选择性地更新有变化的本体部分。

此外如 OntoLT[38] 和 OntoBuilder[39] 等工具也在不断改进中,它们共同推动着 OL 技术的发展。

4　基于模式标注(Pattern - based Annotation)的方法

与前 3 种方法相比,基于模式标注的知识抽取更加注重利用自然语言分析技术。基于模式标注的知识抽取可分为两种类型[40]:一种通过模式的自动发现,进而实现对相关内容的标注;另一种通过人工定义的模式实现内容标注。

基于模式自动发现的模式标注通常遵从 Sergey Brin 提出的反复迭代的模式关系扩展(DIPRE - Dual Iterative Pattern Relation Expansion)方法[41]。

Sergey Brin 以从 Web 上抽取图书作者、题名(Author, Title)对的例子说明这一方法。首先,Brin 利用小规模的(Author, Title)对作为种子集(在实际例子中,仅用了 5 本书的作者和题名对),然后从 Web 上查找这 5 本书所出现的所有实例,从这些实例中,系统识别出描述这 5 本书的各种模式,根据这些模式到 Web 上查找更多新的图书,其后进一步利用这些新图书,查找这些新图书出现的实例,生成更多新的模式,基于此又可利用这些新模式查找新的图书,如此反复迭代,直到从 Web 上识别出大量图书和这些图书的模式。Brin

利用 Python 实现了这一方法，从 5 本图书实例开始，在几乎不需要人工干预的情况下，从 Web 网页中获得了 346 种图书模式，高质量地识别出 15 257 本图书实例。

Armadillo[42] 是基于模式自动发现的另一个系统。它可以从不同数据来源抽取特定领域的标注内容，并将其集成到一个仓储中，形成知识库。它的一个实际应用是挖掘计算机科学系的网站，从中抽取出谁为哪个系工作，其人名、职位、主页、E－mail 地址、电话、一些个人数据，以及这个人发表的论文列表[43]。

除自动发现模式外，基于人工定义的模式标注也是当前知识抽取的重要方法之一。这种方法的代表性系统有 C－PANKOW（及它的前身 PANKOW）[44,45]。

C－PANKOW 被认为是上下文驱动、利用 Web 知识进行基于模式的标注的系统。该系统具有两个特点：

（1）利用无监督的、基于语言分析的模式来识别实例及实例间关系，并将抽取的实例及关系归入到指定的本体中；

（2）将 Web 作为最大的语料库，通过 Google API 计算具有歧义的实例类型。例如出现在文档中的词“Niger”，它可能被标注为一个国家、一个州、一条河或一个地区。C－PANKOW 通过 Google API 计算 Google 检索结果中出现“Niger”的上述 4 种类型的文档和需要标注的目标文档的相似性，最终给出 Niger 的所属类别。

C－PANKOW 主要利用以下 3 种模式来实现语义标注。

（1）Hearst Patterns 模式。利用 Hearst 定义的 4 种模式识别和标注 is_a 关系[46]，这 4 种模式分别是：

H1：<CONCEPT>s such as <INSTANCE>

例如，hotels such as Ritz

H2：such <CONCEPT>s as <INSTANCE>

例如，such hotels as Hilton

H3：<CONCEPT>s，（especially | including） <INSTANCE>

例如，presidents，especially George Washington

H4：<INSTANCE>（and | or） other <CONCEPT>s

例如，the Eiffel Tower and other sights in Paris

（2）定义模式。通过定冠词 the 识别专有名词。C－PANKOW 主要利用以下两种模式：

DEFINITE1：the <INSTANCE> <CONCEPT>

例如，the Hilton hotel

DEFINITE2：the <CONCEPT> <INSTANCE>

例如，the hotel Hilton

（3）同格和连系模式。同格和连系模式分别如下：

APPOSITION：<INSTANCE>，a <CONCEPT>

例如，Excelsior，a hotel in the center of Nancy

COPULA：<INSTANCE> is a <CONCEPT>

例如，The Excelsior is a hotel in the center of Nancy

C－PANKOW（PANKOW）目前已被集成到 OntoMat[47] 和 Magpie[48] 中。除 PANKOW 外，Ontea[49,50] 也是一个基于人工模式实现内容标注的知识抽取系统。

5　语义标注（Semantic Annotation）方法

语义标注除利用自然语言的语法模式和规则外，更重要的是对语义内容的挖掘。在各种文献中，语义标注有多种不同表达方式，如 Semantic Annotation，Semantic Tag，Semantic Markup，Semantically Interlink 等。按照 Atanas Kiryakov 等人的定义，语义标注是为文档中实体提供与它们相关语义描述的过程[51]。Steffen Staab[52] 则更具体地认为，与“不受约束的元数据生成”不一样，语义标注需要实现以下 4 种语义关系的建立：

（1）要唯一标识标注对象，相同的对象用同一标识；

（2）要构建对象和类型的关系，说明标注对象的类别；

（3）要构建对象和属性的关系，说明对象有哪些属性，各自属性值是什么；

（4）要构建对象和对象间的关系。因此他认为，语义标注需要构建语义标注的知识库。

Ontotext Lab 的 KIM 系统是大规模自动语义标注方法应用的代表。KIM 的开发者认为语义标注是命名实体识别和标注两个过程的总和[53]。为了实现语义标注，必须满足以下几个基本条件：

（1）一个定义实体类型的 Ontology（至少需要一个分类表），通过它可以将某些实体和相应类别进行关联；

（2）为每个实体指定一个唯一标识，通过它可以区分不同的实体，同时可以实现实体和语义描述的关联；

（3）需要一个知识库存储实体描述。

基于上述考虑，KIM 认为正式的知识资源建设是语义标注的一个重要环节。

KIM 的知识资源包括 KIM Ontology 和 KIM 世界知识库[54]。目前 KIM Ontology 以 SEKT 的 PROTON 为基础，大约包含 250 个类和 100 个属性，此外，KIM Ontology 还包括 KIM System Ontology 和 KIM Lexical Ontolo-

gy,这两个 Ontology 都是 KIM 在语义标注过程中,对系统功能和语词识别描述的 Ontology。为语义标注过程提供背景知识环境。KIM 世界知识库中预装了大约 900 000条实例描述,主要是人名、地名和组织等一些基本实例,其中包括有 602 585 条人名实例,239 046条组织实例和 50 163 条地名实例。为了让 KIM 发现和标识出不在知识库中的新实体和关系, KIM 知识库还提供了一系列词汇资源,如组织的前后缀、人的尊称、时间格式等,这些都可用于语义标注过程中。KIM 利用基于 Sesame 的 OWLIM 仓储存储这些知识资源,以支持快速大规模的语义标注。

KIM 语义标注本质上是根据 PROTON Ontology 识别和组织存储命名实体的过程。自动标注过程中,发现文献中已标识过的命名实体时,系统将给出这一实体的类型,并将它和知识库中已存在的实例相关联;而对于新的、从未被标识过的实体,系统将在知识库中为其分配一个新的唯一标识并将其存入知识库。

从过程上看,KIM 的语义标注与传统的信息抽取相比有以下特点:

(1)应用基于知识库的语义辞典;

(2)模式匹配语法应用 Ontology 和上下文语义环境;

(3)利用语义概念实现共指消解,如能够通过对北京的语义描述(如别名),判定 Beijing 和 Peking 为同一个城市;

(4)利用知识库实现语义消歧;

(5)所有标注实体都通过它们的类型与 Ontology 关联,通过唯一标识存储在知识库中,并通过它们之间的关系识别建立实体间的关系。

除 KIM 外,类似的语义标注系统还有如 MnM[17]、Artequakt[55]等。

6　基于 Ontology 的信息抽取(OBIE)方法

OBIE 可以认为是当前语义标注研究的一种主流方法。除被称为 OBIE 之外,也有人称其为基于本体的标注(Ontology - based Annotation)和基于本体的语义标注(Ontology - based Semantic Annotation)。

传统信息抽取系统多采用扁平结构组织知识,基于词表、规则或机器学习的方法来抽取文本中的实体。实践证明传统信息抽取系统在关系抽取、歧义消解、可移植性等方面能力十分有限。Embley 提出基于 Ontology 的信息抽取(OBIE)方法[56],希望以这种新的知识描述方式解决传统信息抽取中的难点问题。

OBIE 是上一节中语义标注的进一步发展,它不但要将抽取出的内容纳入到知识库中,还要求在抽取过程中一直得到 Ontology 的支持。OBIE 通过 Ontology 定义的类、属性、层次结构抽取非结构化或半结构化文本中对应的实例,进行歧义消解,进而识别文本中的实体及关系,将结果存储于对应的 Ontology 中。

欧盟 Musing(MUlti - industry, Semantic - based Next Generation Business INtelliGence)[57]是 OBIE 系统的典型代表。Musing 设计了适用于商业领域的 Ontology,并采用 GATE(General Architecture for Text Engineering)[58]作为抽取平台,抽取的准确率较高。

Musing 知识抽取系统的基本思路是:

(1)由领域专家扩充 PROTON 上层本体,定义商业领域的 Ontology,该 Ontology 包含商业领域的类层次结构、关系和属性;

(2)确定好大量用于 OBIE 的信息源(除一些固定合作方提供的数据外,还监测大量商业网站,如 Yahoo! Finance 等);

(3)定期将这些信息源的数据抓取到本地并存储在 Musing 的文档数据库中;

(4)利用 GATE,基于词表和规则从文档中抽取出实体和关系,并通过聚类算法对跨文档的实体和关系进行歧义消解;

(5)用 Musing 特有的 Ontology Mapping 组件把这些实例映射到 Ontology 类和属性中;

(6)采用 RDF statement 生成组件,将实例自动写入 Ontology;

(7)采用有效的数据结构将已获得的实例存储为结构化形式,构建知识库以便于在以后的应用中查询和推理。

面向不同的应用领域和设计目标,OBIE 系统有不同的设计角度,系统实现的技术方法也各不相同,如 IBM 的 Semtag[59]属于基于实例的 OBIE 系统,主要利用 Ontology 的实例实现实体和关系的抽取。该系统并不试图运用规则发现新实例,也不对知识库进行扩充,其目标是抽取的准确率和效率。系统的实现关键是逻辑正确的 Ontology 以及精确实例的支持。该系统适用

于大规模、粗粒度的信息抽取。McDowell 和 Cafarella 等人开发的自动信息抽取系统 OntoSyphon[60]，及由 B. Yildiz 和 S. Miksch 等人开发的 OntoX 系统[61]属于 Ontology 驱动的信息抽取，是从系统可移植性的角度设计的。系统以 Ontology 为起点和核心，在没有人工干预和机器训练的情况下，从 Web 或大量文档集中进行关键词检索，抽取实例自动生成知识库。系统能随着 Ontology 的改变而自动适用于不同领域。

7　基于受控语言的信息抽取(CLIE)方法

与前面 6 种知识抽取技术方法相比，Adam Funk 等提出的基于受控语言的信息抽取（Controlled Language Information Extraction，CLIE）[62]是一种很特殊的技术方法。它以某些受控语言撰写的文本为处理对象，从这些受控语言的文本中构建 Ontology。它可以降低 Ontology 构建的门槛，提高 Ontology 构建效率。

已有的本体构建工具如 Protégé 等，需要用户掌握复杂的知识组织标准，熟悉本体编辑工具的专业知识。这些要求增加了人们管理知识的难度。CLIE 可以简化知识管理中创建结构化数据的过程，增强用户创建、修改和利用存储已有仓储库中知识的能力。CLIE 主要思想是将 CL 与自然语言处理相结合，利用语法规则从符合受控规则的文本中自动抽取类、实例、属性等元数据，进而构建本体的一种新方法。其中，CL 是经过人工定义，在词汇、句法和文体等方面受到控制，仅包含一定量与特定任务相关的词汇条目和语法规则的自然语言子集[62]。

由于 CLIE 最终目的是构建本体，而本体包括类、关系、实例、属性等要素，因此 CLIE 过程需要实现的方法包括：定义新类、创建类之间的层级关系、定义对象和数据类型的属性、创建实例、创建实例的属性值。为实现这些目标，CLIE 构建了管道式流程，如图 2 所示：

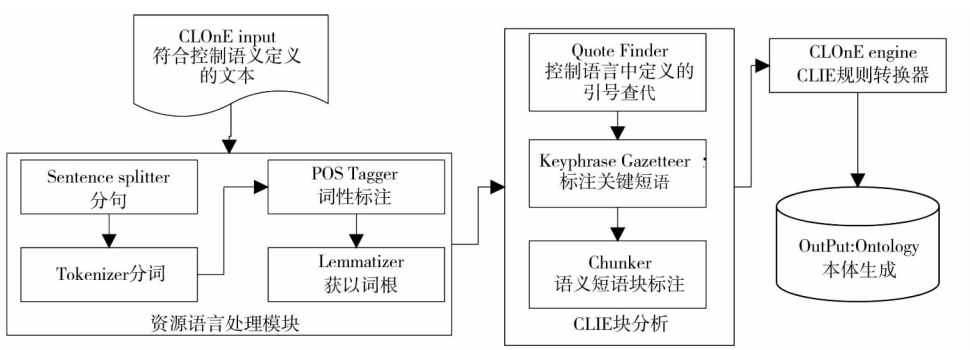

图 2　CLIE 管道流程图[62]

在实现中，CLIE 被划分为两个独立的部分：语言接口 CLOnE（Controlled Language for Ontology Editing）和应用接口 CLIE 组件。

CLOnE 建立在已有的机器翻译和应用 CL 表达知识基础上，借助定义的语法规则、词汇等，规范用户的输入文本。CL 中包含的关键词和类名是 CLIE 词表中对短语进行标注的重要依据，是用来推理词汇间关系的重要保证。

CLIE 组件是基于 GATE 级联有限状态转换器构建的自然语言处理器。处理过程中，CLIE 组件将根据受控语言的语法规则等判断输入文本的有效性，若有效则接收并进入解析过程；若无效则拒绝接收，并提示该文本需要修改的语法。在解析过程中，CLIE 首先选用自然语言处理提供的分句、分词、词性标注和取词根等操作实现文本预处理，其后根据词性、分词等标注，确定命名短语块、分隔符、前置介词和结束标记等。获得确定的句群后，通过 Keyphrase 辞典标注出能反应类关系的短语部分。Chunker 转换器规则中，规则一侧的类正则表达式表示句子模式，当待标注句子与此模式匹配时，通过规则另一侧实现句子语义向本体的转换，从而实现本体构建。

目前，CLIE 得到较多应用，如英国 EPSRC 资助的 Poleazy 项目利用 CLIE[62]为编辑 IT 版权政策本体提供受控自然语言接口。该项目涉及 CL 扩展、本体与 CLOnE 互生成的循环信息流。CLIE 与 Lion[63]项目合作，进一步评测了 CLOnE 对各案例的适应性。

8　结　语

通过对上述典型知识抽取系统的分析,可以发现,知识抽取是在信息抽取的基础之上更加深入地发现文献中隐含知识的过程。总体而言,知识抽取表现出以下 5 个特点:

(1)知识抽取强调语义的抽取。抽取出的内容是有一定意义的、能被其它上下文所解释的语义知识片段(如概念及概念间的关系等)。

(2)知识抽取普遍将机器学习技术和自然语言分析技术相结合。与传统的基于学习或规则的信息抽取不同,由于面对更为复杂的任务,很多知识抽取的系统都采用机器学习技术和自然语言分析技术相结合的方法。

(3)知识抽取需要 Ontology 的支持。Ontology 是知识抽取不可或缺的组件。在知识抽取前,Ontology 定义需要抽取的知识类型;命名实体识别过程中,Ontology 除了能够起到词表和辞典的辅助标识作用外,还可为知识抽取提供推理机制;在语义标注中,Ontology 可以对抽取结果进行语义识别和消除歧义;处理抽取结果,抽取结果被关联到 Ontology 中,形成知识库。

(4)知识抽取关注实体间关系的识别和抽取。知识抽取除了要识别出命名实体的类型外,还需要识别出这一命名实体与其它命名实体之间的各种关系,通过关系将识别出来的新实体纳入到相应的知识库之中。

(5)知识抽取的结果为知识库建设提供了内容。根据预先定义的 Ontology 框架,知识抽取系统从一系列文献中抽取出相应实体和关系,并将这些文献和抽取出的实体和关系组织到知识库中,实现本体填充(Ontology Population)。所建设的知识库是进一步实现数据挖掘、知识发现的基础。

知识抽取的技术方法还在不断地完善和丰富中,自适应的信息抽取、开放信息抽取、OBIE、CLIE 等方法的提出对知识抽取技术的发展做出了有益尝试,而机器学习和自然语言分析两大技术思路的相互融合已经成为知识抽取技术发展的主流趋势。随着知识抽取技术方法的不断完善,知识抽取必将更加深远地影响到语义 Web、知识工程、领域描绘、趋势分析、主题发现、舆情监测、自动问答等诸多与图书情报服务密切相关的领域。

参考文献:

[1] Special Session on Signal Processing Techniques for Knowledge Extraction and Information Fusion in Frame of KES2006 [EB/OL]. [2008 - 06 - 01]. http://www.bsp.brain.riken.jp/kes2006session/.

[2] X - Media Project [EB/OL]. [2008 - 06 - 01]. http://www.x - media - project.org.

[3] K - space Project, Knowledge Space of Semantic Inference of Automatic Annotation and Retrieval of Multimedia Content [EB/OL]. [2008 - 06 - 01]. http://kspace.qmul.net:8080/kspace/index.jsp.

[4] Geoffrey I. Webb. Discovering Significant Patterns [J]. *Machine Learning*, 2007, 68(1):1 - 33.

[5] Alani H, Kim S, Millard D E, Weal M J, Lewis P H, Hall W, Shadbolt N R. Automatic Extraction of Knowledge from Web Documents [C]. In: 2*nd International Semantic Web Conference - Workshop on Human Language Technology for the Semantic Web abd Web Services*, October 20 - 23, 2003, Sanibel Island, Florida, USA.

[6] Martin Rajman, Romaric Besancon. Text Mining - Knowledge Extraction from Unstructured Textual Data [C/OL]. [2008 - 06 - 01]. http://liawww.epfl.ch/Publications/Archive/RajmanBesancon98a.pdf.

[7] AKT Project[EB/OL]. [2008 - 06 - 01]. http://www.aktors.org/akt/.

[8] CLEF: Clinical e - Science Framework [EB/OL]. [2008 - 06 - 01]. http://www.clinical - escience.org/.

[9] SEKT Project [EB/OL]. [2008 - 06 - 01]. http://www.sekt - project.com/.

[10] Dot.Kom Project [EB/OL]. [2008 - 06 - 01]. http://nlp.shef.ac.uk/dot.kom/.

[11] DELOS Project [EB/OL]. [2008 - 06 - 01]. http://www.delos.info/.

[12] OpenKnowledge [EB/OL]. [2008 - 06 - 01]. http://openk.org/.

[13] KnowItAll [EB/OL]. [2008 - 06 - 01]. http://www.cs.washington.edu/research/knowitall/.

[14] Project HALO [EB/OL]. [2008 - 06 - 01]. http://www.projecthalo.com/.

[15] Rapid Knowledge Formation Project [EB/OL]. [2008 - 06 - 01]. http://projects.teknowledge.com/RKF/.

[16] Knowledge Extraction from Document Collections [EB/OL]. [2008 - 06 - 01]. http://www.parc.com/research/projects/knowledge_extraction/.

[17] Vargas - Vera M, Motta E, Domingue J, Lanzoni M, Stutt A, Ciravegna F. MnM: Ontology Driven Semi - Automatic and Automatic

Support for Semantic Markup [C]. In: *The* 13*th International Conference on Knowledge Engineering and Management* (*EKAW* 2002). Springer Verlag Heidelberg, 2002.

[18] KIM Platform [EB/OL]. [2008 - 06 - 01]. http://www.ontotext.com/kim/index.html.

[19] ArtEquAKT [EB/OL]. [2008 - 06 - 01]. http://www.artequakt.ecs.soton.ac.uk/.

[20] Text2Onto[EB/OL]. [2008 - 06 - 01]. http://onto - ware.org/projects/text2onto/.

[21] PowerMagpie [EB/OL]. [2008 - 06 - 01]. http://powermagpie.open.ac.uk/.

[22] Amilcare [EB/OL]. [2008 - 06 - 01]. http://nlp.shef.ac.uk/amilcare/.

[23] Ciravegna F. Adaptive Information Extraction from Text by Rule Induction and Generalisation [C]. In: *Proceedings of* 17*th International Joint Conference on Artificial Intelligence* (*IJCAI* 2001), Seattle. 2001.

[24] Melita[EB/OL]. [2008 - 06 - 01]. http://nlp.shef.ac.uk/melita/.

[25] Ciravegna F, Dingliand A, Petrelli D. Active Document Enrichment Using Adaptive Information Extraction from Text[C]. *In*: 1*st International Semantic Web Conference* (*ISWC* 2002), June 9 - 12th, 2002, Sardinia, Italia.

[26] Handschuh S, Staab S, Ciravegna F. S - CREAM - Semi - automatic CREAtion of Metadata [C]. In: *Processdings of the European Conference on Knowledge Acquisition and Management*(*EKAW*02), Springer, 2002.

[27] Banko M, Cafarella M J, Soderland S, Broadhead M, Etzioni O. Open Information Extraction from the Web [C]. In: *Proceedings of the* 20*th International Joint Conference on Artificial Intelligence* (*IJCAI* 2007).

[28] TextRunner [EB/OL]. [2008 - 06 - 01]. http://www.cs.washington.edu/research/textrunner/.

[29] Yates A. Information Extraction from the Web: Techniques and Applications[D/OL]. [2008 - 06 - 01]. http://turing.cs.washington.edu/papers/yates_dissertation.pdf.

[30] Klein D, Manning C D. Accurate Unlexicalized parsing [C/OL]. In: *Proceedings of the ACL*, 2003. [2008 - 06 - 01]. http://www.cs.berkeley.edu/ ~ klein/papers/unlexicalized - parsing.pdf.

[31] The OpenNLP Home [EB/OL]. [2008 - 06 - 01]. http://opennlp.sourceforge.net/projects.html

[32] Sunción Gómez - Pérez A, Manzano - Macho D. A survey of ontology learning methods and Techniques[R]. OntoWeb Deliverable D1.5, 2003,6.

[33] Cimiano P. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications[M]. Springer US,2006:19 - 34.

[34] Buitelaar P, Cimiano P, Grobelnik M. Ontology Learning from Text [C]. In: *the ECML/PKDD* 2005 *Workshop on*: *Knowledge Discovery and Ontologies*, 2005.

[35] Buitelaar P, Cimiano P, Magnini B. Ontology Learning from Text: An Overview[M]. IOS Press, 2003.

[36] KAON - The KArlsruhe ONtology and Semantic Web tool suite [EB/OL]. [2008 - 06 - 01]. http://kaon.semanticweb.org/.

[37] Cimiano P, Völker J. Text2Onto - A Framework for Ontology Learning and Date - driven Change Discovery[C]. In: *Proceedings of NLDB*05, June 2005.

[38] OntoLT [EB/OL]. [2008 - 06 - 01]. http://olp.dfki.de/OntoLT/OntoLT.htm.

[39] OntoBuilder [EB/OL]. [2008 - 06 - 01]. http://iew3.technion.ac.il/OntoBuilder/.

[40] Reeve L, Han H. Survey of Semantic Annotation Platforms[C]. In: *Proceedings of the* 2005 *ACM symposium on Applied computing*, New York: ACM Press, 2005:1634 - 1638.

[41] Sergey Brin, Extracting Patterns and Relations from the World Wide Web[C]. In: *WebDB Workshop at* 6*th International Conference on Extending Database Technology*, 1998.

[42] Armadillo [EB/OL]. [2008 - 06 - 01]. http://www.dcs.shef.ac.uk/ ~ sam/armadillo.html.

[43] Ciravegna F, Chapman S, Dingli A, Wilks Y. Learning to Harvest Information for the Semantic Web[C]. In: *Proceedings of the* 1*st European Semantic Web Symposium*, Heraklion, Greece, May 10 - 12, 2004.

[44] Cimiano P, Handschuh S, Staab S. Towards the Self - Annotating Web [C]. In: *Proceedings of the* 13*th WWW Conference*, ACM, New York, 2004:462 - 471.

[45] Cimiano P, Ladwig G, Staab S. Gimme' the Context: Context - driven Automatic Semantic Annotation with C - PANKOW [C]. In: *Proceedings of the* 14*th International Conference on World Wide Web*, New York: ACM Press, 2005:332 - 341.

[46] Hearst M A. Automatic Acquisition of Hyponyms from Large Text Corpora. [C/OL]. In: *Proceedings of the* 14*th International Conference on Computational Linguistics*. [2008 - 06 - 01]. http://acl.ldc.upenn.edu/C/C92/C92 - 2082.pdf.

[47] OntoMat - Annotizer [EB/OL]. [2008 - 06 - 01]. http://annotation.semanticweb.org/ontomat/index.html.

[48] Dzbor M, Motta E. Study on Integrating Semantic Applications with Magpie[C]. In: 15*th Conf. on AI Methodology*, *Systems & Applications* (*AIMSA*), Varna, Bulgaria. 2006.

[49] Laclavik M, Seleng M, Babik M. OnTeA: Semi - automatic Ontology Based Text Annotation Method[C]. *ITAT* 2006, *NAZOU Workshop*, 26.9 - 1.10. 2006, Chata Kosodrevina, Bystrá dolina,

Nízke Tatry, 2006.

[50] Laclavik M, Gatial E, Balogh Z, Habala O, Nguyen G, Hluchy L. Semantic Annotation Based on Regular Expressions[C]. ITAT 2005, 20 – 25 September 2005, Hotel Akademik, Rackova dolina, In: *Proceedings of ITAT* 2005 *Information Technologies – Applications and Theory*, Peter Vojtas (Ed.), Prirodovedecka Fakulta Univerzity Pavla Jozefa Safarika v Kosiciach. Slovakia, September 2005:305 – 306.

[51] Kiryakov A, Popov B, Terziev I, Manov D, Ognyanoff D. Semantic Annotation, Indexing, and Retrieval [C]. *Elsevier's Journal of Web Semantics*, 2003:484 – 499.

[52] Staab S, Maedche A, Handschuh S. An Annotation Framework for the Semantic Web [C]. In: *Proceedings of the First Workshop on Multimedia Annotation*, Tokyo, Japan, January 30 – 31, 2001.

[53] Popov B, Kiryakov A, Ognyanoff D, Manov D, Kirilov A, Goranov M. Towards Semantic Web Information Extraction[C]. In: *Human Language Technologies Workshop at the* 2*nd International Semantic Web Conference* (*ISWC*2003), October 20, 2003, Florida, USA.

[54] Manov D, Popov B. Massive Automatic Annotation[EB/OL]. [2008 – 06 – 01]. http://www.sekt – project.org/rd/deliverables/wp02/sekt – d – 2 – 6 – 1 – Massive% 20Automatic% 20Annotation% 20V1.pdf.

[55] Alami H, Kim S, Millard D E, et al. Automatic Ontology – based Knowledge Extraction from Web Document[J]. *IEEE Intelligent Systems*, 2003,18(1):14 – 21.

[56] Embley D W, Campbell D M, Smith R D, Liddle S W. Ontology – Based Extraction and Structuring of Information from Data – Rich Unstructured Documents[EB/OL]. [2008 – 06 – 01]. http://pages.cs.wisc.edu/ ~ smithr/pubs/cikm98.pdf.

[57] Saggion H, Funk A, Maynard D, Bontcheva K. Ontology – based Information Extraction for Business Intelligence [EB/OL]. [2008 – 06 – 01]. http://iswc2007.semanticweb.org/papers/837.pdf.

[58] GATE. General Architecture for Text Engineering [EB/OL]. [2008 – 06 – 01]. http://gate.ac.uk/.

[59] Dill S, Eiron N, Gibson D, et al. SemTag and Seeker:Bootstrapping the Semantic Web Via Automated Semantic Annotation [C]. In: *Proc. of the* 12*th Intl. WWW Conf.* 2003. Hungary: ACM Press.

[60] Luke K. McDowell, M. C. Ontology – driven Information Extraction with OntoSyphon[EB/OL]. The 5th International Semantic Web Conference(2006). [2008 – 06 – 01]. http://turing.cs.washington.edu/papers/iswc2006McDowell – final.pdf.

[61] Yildiz B, Miksch S. OntoX – A Method for Ontology – Driven Information Extraction[C]. Computational Science and Its Applications – ICCSA 2007, Springer – Verlag, LNCS 4707.

[62] Funk A, Davis B, Tablan V, Bontcheva K, Cunningham H. Controlled Language IE Components Version 2. SEKT Project Deliverable D2.2.2. January 2007 [EB/OL]. [2008 – 06 – 01]. http://gate.ac.uk/projects/sekt/deliv2 – 2 – 2.pdf.

[63] Líon Project[EB/OL]. [2008 – 06 – 11]. http://www.deri.ie/research/projects/.

（作者 E – mail:zhangzhx@mail.las.ac.cn）