基于Ontology的信息抽取（OBIE）技术方法分析*

洪娜[1,2]　张智雄[1]　刘建华[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）

[摘要]本文通过对国内外OBIE理论和OBIE系统的分析，比较了OBIE技术与传统IE技术的主要区别，归纳了四种主要的技术方法，分别是基于实例的OBIE，基于规则的OBIE，基于机器学习的OBIE和Ontology驱动的OBIE，并用案例对各种技术方法做了阐释，最后总结了OBIE研究和系统开发中仍然存在的难点问题。

[关键词]OBIE Ontology 信息抽取

[分类号]G250.76

Technical Methods Analysis of Ontology-based Information Extraction

Hong Na[1,2]　Zhang Zhixiong[1]　Liu Jianhua[1,2]

[1](*Library of Chinese Academy of Sciences, Beijing 100190, China*)

[2] (*Graduation School of Chinese Academy of Sciences, Beijing 100049, China*)

[Abstract] After analyzing the theories and system of OBIE, this paper compares the technology of OBIE and traditional IE, and concludes four main technical methods including instance-based OBIE, ruled-based OBIE, machine learning-based OBIE and Ontology-driven information extraction, then explains these technical methods by practical examples. Finally this paper sums up some problems still exist in OBIE study and system development nowadays.

[Keywords] OBIE, Ontology, Ontology-based information extraction

1 引　言

传统的IE系统大多采用平板结构组织知识，采用基于词表、规则或机器学习的方法来抽取文本中的实体，但实践证明传统IE系统在关系抽取、歧义消解、可移植性等方面的能力十分有限。Embley提出了一种利用Ontology抽取非结构化文本中的信息内容的方法[1]，希望以一种新的知识组织方式解决传统IE中难点问题。目前，越来越多的IE系统将Ontology用于信息抽取任务，作为传统IE系统的改进，成为当前的一个热点研究领域，OBIE（Ontology-based information extraction）即基于Ontology的信息抽取，是一个包含了多种自然语言处理技术的综合研究领域，简单来说，OBIE是通过Ontology描述的类、属性、层次结构抽取非结构化文本或半结构化文本中对应的实例并对其进行歧义消解，进而识别文本中的实体及其关系，并将其存储于对应的Ontology语义结构中的方法。

OBIE技术的主要优势在于：

（1）相对于传统IE系统采用平板结构组织知识，OBIE采用Ontology结构组织知识，它可以有效的定义实体和关系。

（2）传统IE不能很好的处理歧义消解问题，必须先识别出一篇文章中的所有实体，才能对其进行歧义消解，实现难度较大；OBIE系统可以将本文中的实例直接与Ontology中

* 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”（课题编号：05BTQ006）的研究成果之一。

的类、属性关联起来，从而有效的处理首语重复和指代消解等问题。

（3）基于规则的传统 IE 方法，在处理关系的抽取时需要编写的规则非常复杂，而基于机器学习的传统 IE 方法在训练实体时有较好的效果，但不能很好的进行关系的训练。而 OBIE 系统可以较好的解决知识抽取中的关系抽取、事件抽取等难点问题。

（4）OBIE 系统的抽取结果和定义的 Ontology 密切相关，而 Ontology 和知识库可以通过专家修订等方式不断的完善，这一点与传统 IE 相比专业性更强，应用效果也会更理想。

（5）OBIE 系统的应用领域可以随着 Ontology 的不同而不同，只需为每一个 Ontology 进行少量的语料标注和训练。而不像传统 IE 系统移植到新领域时，需要改动大量的规则或重新训练大量的数据。

（6）OBIE 系统随着 Ontology 的更新，自动抽取的信息总是与 Ontology 同步更新，相比传统 IE 在更新词表和规则时需要人工进行大量的工作，OBIE 可以避免大量的人工劳动。

（7）传统IE抽取的结果简单存储于关系数据库中，不能体现知识积累和丰富的过程，而OBIE和Ontology知识库的丰富化是互相支持、共同进化的统一过程。

目前，OBIE 相关理论和技术的研究受到了国内外信息抽取领域的关注，因为 OBIE 系统可以将本文中的实例直接与 Ontology 中的类、属性关联起来，对知识抽取中的关系抽取、事件抽取等难点问题的解决有着重要的意义。OBIE 促进知识库实例的不断丰富，这对知识的进一步分析和重用有重要的作用。另外，OBIE 系统并不试图对文本的语义进行全面深刻的自然语言解析，OBIE 系统的应用通常都是面向特定领域的特定的任务的，这一点和知识抽取任务的面向领域的要求是一致的，因此，针对特定的知识抽取任务做详细规划和任务分解，构建问题求解模型，开发有效、强健的 OBIE 系统是具有可行性的。

2 当前 OBIE 的研究现状

目前，国内外 OBIE 技术的研究已取得了初步的成果。国内从事 OBIE 技术研究的代表机构有北京大学计算语言学研究所，其在 OBIE 领域的研究主要有语料的收集、分类和标注、亚洲语言的自然语言处理和多语言知识库的建设等；哈尔滨工业大学的信息检索实验室，其在 OBIE 领域的研究主要有 Ontology 的自动构建、关系的抽取、事件的抽取和指代消解等。面向特定领域的 OBIE 应用研究也有一些探索，如兰州大学开发出从生物文本中抽取抗艾滋病抑制剂的 OBIE 系统[2]。

国外从事 OBIE 技术和应用研究的机构和项目较多，如谢菲尔德大学的 SEKT 和 hTechSight 项目[3]、OntoText 实验室的 KIM 项目[4]、南安普敦大学的 ArtEquAkt 项目[5]、AIFB 机构的 C-PANKOW 系统[6]、由 McDowell 和 Cafarella 等人开发的自动信息抽取系统 OntoSyphon[7]、由 B. Yildiz 和 S. Miksch 等人开发的 ontoX 系统[8]等，对以上典型 OBIE 系统对比分析，KIM 和 SEKT 的 OBIE 系统采用 GATE（General Architecture for Text Engineering）[9]作为 IE 工具，抽取精准率更高，KIM 的 OBIE 系统是基于规则的 IE 系统，基于 Ontology 结构上采用模式匹配方法，SEKT 的 OBIE 系统是采用机器学习方法开发的 IE 系统。

OBIE 的研究在 Ontology 的构建、实体抽取和知识库的装载等方面已经取得了不少的成果，目前，其研究重点转向关系的抽取、事件的抽取等。如，谢菲尔德大学 SEKT 项目 OBIE 系统的研究最近重点采用机器学习的 IE 方法解决关系抽取的问题，ArtEquAkt 项目的研究目的是利用 OBIE 生成的知识库中的知识应进行自动的自传编写。

目前，OBIE 的研究还面临许多挑战，主要有：

（1）移植性。随着 Ontology 的改变而抽取不同的内容。在没有人工干预和机器训练

的情况下，使机器能随着 Ontology 的改变而自动应用于不同领域。

（2）处理不同文本类型的能力。如结构化、半结构化、自由文本等。

（3）机器学习方法应用在 Ontology 的程度。需要用标注过的小量语料进行训练，想要训练的概念越多，需要标注的数据也就越多。标注数据的成本是很高的。但是 OBIE 的结果又和训练的数据量密切相关。

（4）自动 OBIE 系统的可靠性方面还需进一步增强。

3 当前 OBIE 的主要技术方法分析

通过对国内外多个有代表性的 OBIE 系统的调研和分析，本文归纳了 OBIE 系统的基本任务流程，并在此基础上提出了 OBIE 的四种主要技术方法。通常，OBIE 系统的基本任务流程是：首先预定义一个核心 Ontology，包含特定领域的类和属性，装载基础的 Ontology 实例，然后从文档中抽取出对应于 Ontology 实例的实体或关系，把新实例装载入 Ontology 或作为 Ontology 类或属性本身用来不断完善核心 Ontology，完善后的 Ontology 再重新用于抽取，OBIE 系统还应采用有效的数据结构将已获得的实例存储为结构化的形式，即构建知识库便于以后的查询、推理和应用[10]。基于以上 OBIE 的基本任务流程，由于应用领域和设计目标的不同，OBIE 系统有不同的设计角度，系统实现的技术方法又各不相同，笔者将主要的 OBIE 的技术方法划分为以下四种。

3.1 基于实例的 OBIE

基于实例的 OBIE 系统主要利用 Ontology 中的实例进行实体和关系的抽取，并不试图运用规则来发现新实例，不对知识库进行丰富，其目标是抽取的精准率和效率。基于实例的 OBIE 实现的关键是逻辑正确的 Ontology 及其精确实例的支持，适合应用于大规模、粗粒度的信息抽取。该种 OBIE 系统的代表是 IBM 的 Semtag[11]和 KMI 的 Megpie[12]，以 Semtag 为例介绍基于实例的 OBIE 系统的流程。

SemTag 是 IBM 开发的语义标注平台，其工作流如图 1 所示，SemTag 选择 TAP 作为其 Ontology 进行语义标注，SemTag 采用 Seeker 进行开发，基于 Annotea 标注框架对 web 文本进行大规模自动语义标注。SemTag 使用词袋法（bag-of-words）进行歧义消解，而 TAP 中定义的实体命名方式比较简单，这也使 SemTag 的歧义消解问题比较好解决，系统采用 Taxonomy Based Disambiguation (TBD) 算法实现。为了正确的将识别出来的对象与 TAP Ontology 中的类进行对应，SemTag 采用基于向量空间模型的算法进行分类。SemTag 采用高性能的并行计算方式处理文档，其单个节点的效率可达到每秒 200 个文档的语义标注。但是 TAP 过于简单的负面的作用是识别的召回率不高，只包含一些常见术语的分类，如音乐、电影和运动等实例共 65K，不能满足细粒度的标注要求。

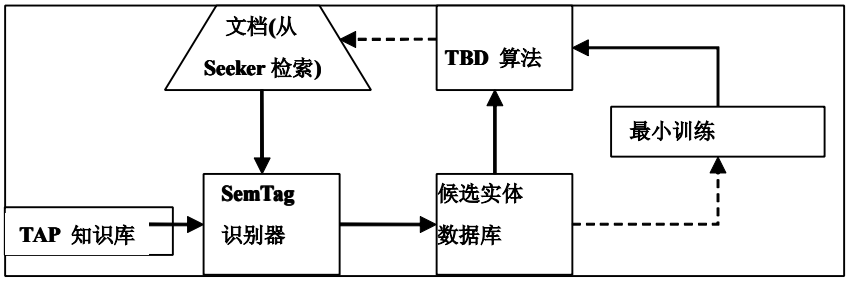

图 1　SemTag 的工作流[13]

3.2 基于规则的 OBIE

基于规则的 OBIE 系统主要除了利用 Ontology 中的实例进行实体和关系的抽取，还利用 Ontology 或语料库构建规则来发现新实例，对知识库进行不断丰富，其目标是抽取的召回率和知识库的动态更新。基于规则的 OBIE 实现的关键是规则的正确编写与新实例的装载，适合应用于精粒度的信息抽取。但此类系统的移植性较差，面向不同的应用需要修改抽取规则，还需要大量训练数据和领域专家的支持。该种 OBIE 系统的代表是 ontotext 语义技术实验室的 KIM，南安普敦大学的 ArtEquAkt 和谢菲尔德大学的 hTechSight，以 KIM 为例介绍基于规则的 OBIE 系统的流程。

KIM（Knowledge & Information Management）是由ontotext语义技术实验室开发的语义标注平台，其工作流如图2所示，KIM选择GATE作为信息抽取的平台，选择KIMO作为其Ontology进行信息抽取，KIM预装了大约900K的实例，主要是人名、地名和组织等一些基本实例，如602585 个人名实例，239046个组织实例和50163个地名实例[1]，这些实例用于基本的信息抽取。抽取的过程要首先用GATE对文本进行词性标记、分句、分词和命名实体识别等基本处理，然后进行Ontology中已有实例的抽取，利用实体排名算法（Entity Ranking algorithm）解决指代消解的问题，指代消解对下一步的关系抽取的准确率和召回率十分重要。此外，KIM基于Ontology对原GATE的规则集进行修改，用来发现文本中的新实体和新关系，进而通过新实例的装载进一步完善Ontology的实例库，实现知识库的动态更新。由于KIM使用固定的本体KIMO，这也相应的限制了KIM的标注范围。

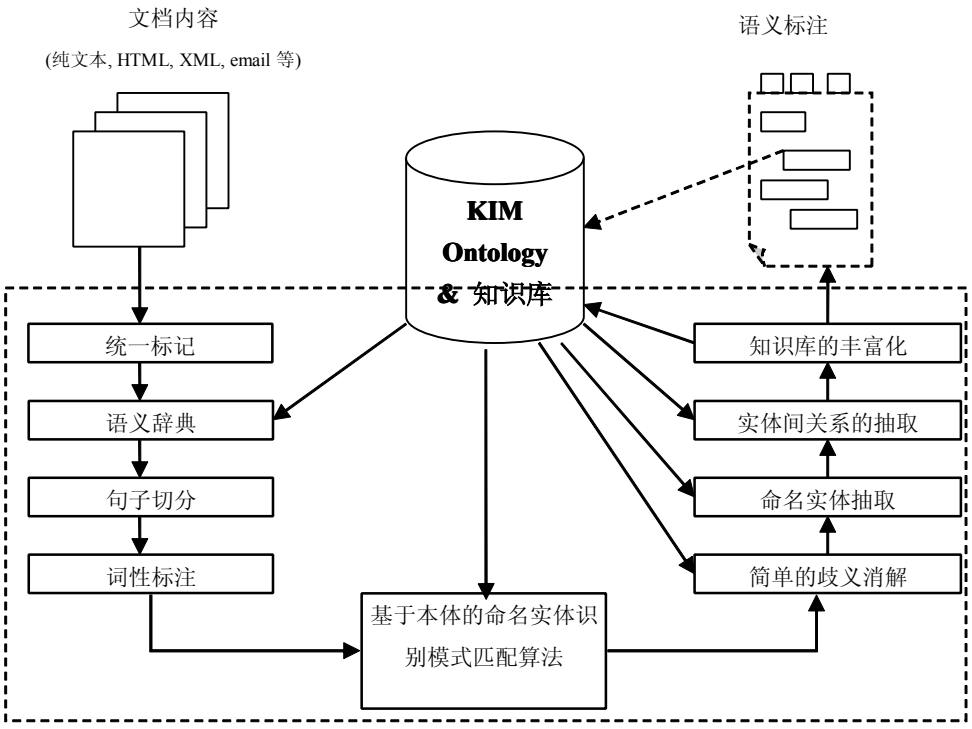

图 2　KIM 的工作流[14]

基于规则的 OBIE 系统的关键点，首先是规则的产生，规则的产生需要通过对 Ontology 的处理，解析出 Ontology 中的抽取规则，必要时可以辅助以手工修正后用作 OBIE 系统的抽取规则。这里特别要解决的问题是：

[1] 该数据摘自 2008 年 3 月 12 日 KIM 网站

（1）基于 Ontology 或语料库的自动规则抽取算法的实现。

（2）规则的 domain 和 range.只有满足 Ontology 限制的规则，才能应用于识别实例。

（3）规则的合理性检验。

另一个关键点是新实例的装载，即如何判断新实体和新关系对应的 Ontology 类和属性以及对它们的分类处理，在此要处理的几个问题是：

（1）对抽取到的新实体和新关系进行定期监测，统计一定周期内出现的候选实例的次数，达到一定阈值的作为装载 Ontology 的新实例对象。

（2）新实例在 Ontology 中有对应的已定义的类和属性，将其作为新实例装载。

（3）新实例在 Ontology 中没有对应的已定义的类和属性，则需要新建 Ontology 中的类和属性。这时还需要考虑新建类和属性在 Ontology 中的层次位置，是十分复杂的问题。

（4）新实例装载之后的 Ontology 一致性检验、矛盾消解和手工修正，以便用于下一个周期的 OBIE。

从以上分析来看，基于规则的 OBIE 系统自身固有的缺点是它对规则的依赖，这也就导致了此类系统的非动态性，当面向不同的应用领域时需要重新针对新的 Ontology 进行领域规则的解析和修改。

3.3 基于机器学习的 OBIE

基于机器学习的 OBIE 系统主要采用机器学习算法来实现 OBIE 任务，其目标是最大程度的实现关系的抽取。基于机器学习的 OBIE 实现的关键是算法在关系识别任务中的正确率和效果，适合应用于精粒度的信息抽取。该种 OBIE 系统的代表是谢菲尔德大学的 SEKT，以 SEKT 为例介绍基于机器学习的 OBIE 系统的流程。

SEKT 是欧盟 6FP（6th Framework programme）联合资助的语义知识技术（Semantic Knowledge Technologies）项目，选择 PROTON[15]作为其 Ontology 进行知识抽取，SEKT 采用 GATE 进行开发，采用研讨会资料和工作相关的语料库作为机器学习的对象。SEKT 认为，机器学习方法可以较好的解决基于 Ontology 的关系抽取问题，机器学习方法的本质就是构建一个分类器，将文本中分析出的多种关系分类对应于 Ontology 的属性实例中，多种机器学习的方法可用于关系抽取，例如，隐马尔可夫链 HMM 算法，环境自由域 CRF 算法，最大熵模型 MEM 算法以及支持向量机 SVM 算法。SEKT 项目认为 SVM 算法是效果最好的关系抽取算法，并在 SVM 和感知器模型上引入不均匀裕度参数（uneven margins parameter）来改进算法性能，称改进后的算法为 SVMUM 和 PAUM[16]。GATE 也为关系的抽取提供工具支持，如 gate.obie API 就包含多种分类器 (TextGarden, WEKA, Maxent, SVM light)，但总的来说，关系的抽取目前还是 OBIE 领域的难点之一。

3.4 Ontology 驱动的 OBIE

严格意义上，以上提到的三种 OBIE 实现方法都是属于文档驱动的 OBIE，IE 的过程依赖于对文本的分析，而 Ontology 驱动的 OBIE 是从可移植性的角度来设计系统，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，使系统能随着 Ontology 的改变而自动适用于不同领域[17]。Ontology 驱动的 OBIE 的实现难度较大，只能适用于部分对语义理解要求不高的场合，该种 OBIE 系统的代表是由 McDowell 和 Cafarella 等人开发的自动信息抽取系统 OntoSyphon，以及由 B. Yildiz 和 S. Miksch 等人开发的 ontoX 系统等，以 OntoSyphon 为例介绍 Ontology 驱动的 OBIE 系统的流程。

OntoSyphon 的目标是从 web 上检索相关文档，抽取实例生成知识库。抽取过程以

Ontology 为起点，利用搜索工具在 web 上做基于关键词（Ontology 的类）的搜索，从 web 上得到大量相关网页和文本，然后计算网页和 Ontology 类的相似度，执行抽取后将实例装载进入知识库[18][19]。OntoSyphon 使用 Ontology 来判断候选实例的相关性，但是它不对文档进行标注。通过对不同领域的实验，证明 OntoSyphon 的可移植性可以达到一般性抽取任务的需求。不同领域 OntoSyphon 的效果如图 3 所示。

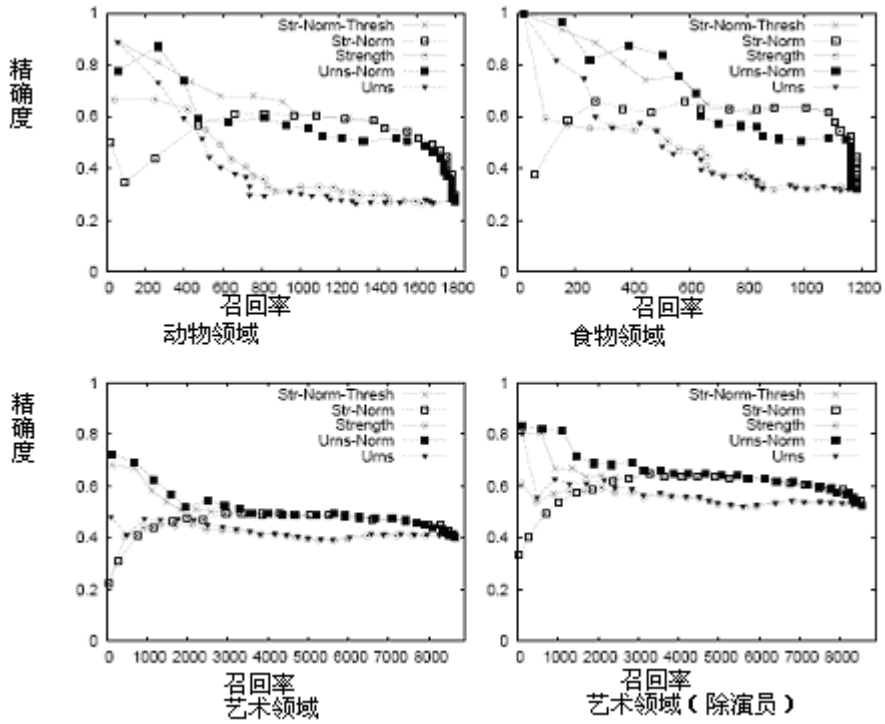

图 3　不同领域 OntoSyphon 的效果[20]

Ontology 驱动的 OBIE 目前并没有十分成熟的实验系统，而且这方面的定义仍有争议，许多场合 Ontology 驱动的 OBIE 和 OBIE 并没有严格的区分。

4 结　语

总体来看，基于实例的 OBIE 在大量实例支持的基础上，易于实现，适合应用于大规模、粗粒度的信息抽取；基于规则的 OBIE 在技术实现上比较复杂，其大规模、精粒度抽取的前提是规则的正确编写和 Ontology 新实例的装载策略;基于机器学习的 OBIE 的关键是机器学习算法的选择和效率；Ontology 驱动的 OBIE 的实现难度较大，基于可移植性的要求，此类系统主要适用于对语义理解要求不高的场合。

目前，OBIE 技术的发展并不成熟，其研究仍有多个难点问题待于解决，如关系的抽取、事件的抽取、多语种信息抽取、跨文档共指消解、Ontology 的进化和 OBIE 系统的移植等。这些问题在当今的 OBIE 系统中还不能很好的解决，此外，OBIE 应用的规范辞典、抽取规则和 Ontology 的定义都应随着 OBIE 的应用需要动态的调整和完善。现今仍然缺少采用机器学习技术建立 OBIE 模型的综合方法。基于 OBIE 的研究现状，下一步在其各个问题领域都有待做更深入的研究。

参 考 文 献

[1] David W. Embley, D.M.C., Randy D. Smith, Stephen W. Liddle, Ontology-Based Extraction and Structuring of Information from Data-Rich Unstructured Documents[EB/OL].(1998)[2008-06-12]. http://pages.cs.wisc.edu/~smithr/pubs/cikm98.pdf

[2] Chunyan Zhang, J.D., Ruisheng Zhang,Xiaoliang Fan,Yongna Yuan,Ting Ning. Extracting Information of Anti-AIDS Inhibitor from the Biological Literature Based on Ontology[M]. Lecture Notes in Computer Science.Springer Berlin / Heidelberg,2007, Volume 4613:74-83.

[3] Diana Maynard, Milena Yankova, Niraj Aswani and Hamish Cunningham.Automatic Creation and Monitoring of Semantic Metadata in a Dynamic Knowledge Portal[M]. Lecture Notes in Computer Science.Springer Berlin / Heidelberg,2004,Volume 3192:74-83.

[4] Borislav Popov, Atanas Kiryakov, Angel Kirilov, Dimitar Manov, Damyan Ognyanoff, Miroslav Goranov. KIM - Semantic Annotation Platform. [2008-06-12].http://www.ontotext.com/publications/KIM_SAP_ISWC168.pdf

[5] Harith Alani, S.K., David E. Millard, Mark J. Weal,Wendy Hall, and N.R.S. Paul H. Lewis. Automatic Ontology-Based Knowledge Extraction from Web Documents[M]. IEEE Computer Society, 2003: 1094-7167.

[6] AIFB. TextToOnto - A paper for end users[EB/OL]. [2008-06-12].http://www.mirrorservice.org/sites/download.sourceforge.net/pub/sourceforge/t/te/texttoonto/TextToOntoPaper.pdf.

[7] Luke K. McDowell, M.C. Ontology-driven Information Extraction with OntoSyphon[EB/OL]. The 5th International Semantic Web Conference(2006) [2008-06-05]. http://turing.cs.washington.edu/papers/iswc2006McDowell-final.pdf

[8] Burcu Yildiz and Silvia Miksch.ontoX-A Method for Ontology-Driven Information Extraction[M],Computational Science and Its Applications-ICCSA 2007,Springer-Verlag,LNCS 4707,p.660-673

[9] GATE. General Architecture for Text Engineering[EB/OL].[2008-06-02]. http://gate.ac.uk/.

[10] W3CHINA.ORG. OWL 的存储方法的选择[EB/OL]. [2008-03-02].http://www.ieee.org.cn/dispbbs.asp?BoardID=2&id=20862&replyID=2653&star=1&skin=0.

[11] Dill, E., Gibson, Gruhl, Guha, Jhingran et al. SemTag and Seeker:bootstrapping the semantic web via automated semantic annotation[C].In Proc.of the 12th Intl.WWW Conf.2003.Hungary:ACM Press p.178-186.

[12] DZBOR, M., DOMINGUE, J. B., MOTTA, E. Magpie - towards a semantic web browser. In Proc. of the 2nd Intl. Semantic Web Conf., October 2003, Florida US.[2008-06-01]. http://kmi.open.ac.uk/people/dzbor/public/2003/iswc03-p47-dzbor-domingue-motta.pdf

[13] 同[11]

[14] 同[4]

[15] SEKT. PROTON ONTOLOGY. [2008-06-02].http://proton.semanticweb.org/.

[16] Yaoyong Li, H.C., etc, Ontology-Based Information Extraction (OBIE) v.1,v.2,v.3[EB/OL]. SEKT deliverable, 2006. [2008-06-02]. http://www.sekt-project.com/rd/deliverables/wp02/sekt-d-2-1-1-Ontology-Based%20Information%20Extraction.pdf

[17] Burcu Yildiz, S.M. Motivating Ontology-Driven Information Extraction[EB/OL]. in Proceedings of the International Conference on Semantic Web and Digital Libraries (ICSD-2007). (2007) [2008-06-12]. http://www.donau-uni.ac.at/imperia/md/content/department/ike/ike_publications/2007/refereedconferenceandworkshoparticles/yildiz_2007_icsd_ontolgy_management.pdf

[18] ZHANJUN LI, K.R. Ontology-based design information extraction and retrieval[J]. Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 2007(21)137 - 154.

[19] Zhu, D.J., Uren, Dr. Victoria, Motta, Prof. Enrico. ESpotter: Adaptive Named Entity Recognition for Web Browsing[EB/OL].Knowledge Management Conference (WM2005). (2005)[2008-06-12]. http://kmi.open.ac.uk/people/jianhan/zhuetal_WM.pdf

[20] 同[7]

作者简介

洪娜（1980－），女，中国科学院文献情报中心，在读博士研究生，发文 5 篇。
张智雄（1971－），男，中国科学院文献情报中心，研究馆员，发文 60 余篇。
刘建华（1984－），女，中国科学院文献情报中心，在读硕士研究生，发文6篇。
通讯地址：北京市海淀区中关村北四环西路 33 号，中国科学院文献情报中心，邮编 100190。电子信箱：hongn@mail.las.ac.cn