Computer Science and Application 计算机科学与应用, 2011, 1, 91-96
http://dx.doi.org/10.12677/csa.2011.13019 Published Online December 2011 (http://www.hanspub.org/journal/csa)

An Approach of Knowledge Extraction Restrained by Ontology

Guojie Li*, Dengfeng Xu
Dalian Hengyi Technology Incorporated Company, Dalian
Email: ligj@hengyi.ln.cn
Received: Sep. 1st, 2011; revised: Sep. 25th, 2011; accepted: Oct. 9th, 2011.

Abstract: In terms of knowledge view, an approach of knowledge extraction restrained by ontology is proposed in this paper. At first, translate a domain ontology into a model expressed by Alloy language and use the form of solution space to express the recognized entities and the recognized entity-relations which can be got when applying named entity recognition technology and entity relation extraction technology to coarse text block in turn. And then translate every solution of solution space into an assertion sentence which will be included in Alloy model. Next, reduce solution space by applying Alloy analyzer to the Alloy model. At last, a whole knowledge instance will be obtained.

Keywords: Knowledge Extraction; Ontology; Alloy

一种基于本体约束的知识抽取方法

李国杰*，许登峰
大连恒宜科技有限公司，大连
Email: ligj@hengyi.ln.cn
收稿日期：2011 年 9 月 1 日；修回日期：2011 年 9 月 25 日；录用日期：2011 年 10 月 9 日

摘　要：从知识的角度出发，提出一种基于本体约束的知识抽取方法：将领域本体中蕴含的逻辑信息转换为一个 Alloy 语言表示的模型，将命名实体识别和实体关系抽取的成果映射为解空间，接着将解空间里的每一个解转化为 Alloy 语言表示的断言语句，然后使用 Alloy 分析器来约简解空间，最终得到一个具有明确语义的完整知识实例。

关键词：知识抽取；本体；Alloy

1. 引言

在信息抽取领域，命名实体识别方法可以为实体赋予正确的语义，实体关系抽取则可以明确两个实体间的关系语义。但从知识的角度来看，无论是命名实体识别还是实体关系抽取都属于“小粒度”(知识片段)的抽取，抽取结果并非独立完整的知识实例。命名实体识别和实体关系抽取的成果只有经过知识合成才能成为独立完整的知识实例。但由于知识体系的复杂性和多样性，合成“知识片段”往往是非常困难的。此外，传统的信息抽取主要采用基于规则的方法，这种方法会导致在抽取信息时因无法克服不同的信息项满足同一或具有包含关系抽取规则时无法确定信息项类型的问题。

本体是共享的、规范化的概念模型，是对某一领域中知识结构的系统描述，因此从知识的角度来看，领域本体是知识抽取最有效的工具之一。但从目前来看，基于本体信息抽取和知识抽取的研究工作对领域本体的利用非常有限，且主要集中在一些信息本身格式和信息上下文格式的利用；而领域本体中蕴含的逻辑知识以及推理能力很少被利用，因此这类方法在抽取结构比较复杂的知识时抽取效果就明显下降了。

本文研究的思路是：将领域本体中蕴含的逻辑知识转换为一个 Alloy 语言表示的模型，在完成命名实

体识别和实体关系抽取的基础上，使用约束逻辑求解方法(Alloy 分析器)来对领域内的信息进行更为精确的定位，最终得到一个具有明确语义的知识实例。本文的组织如下，第二节给出两个引例，明确本文要解决的问题；第三节介绍相关的概念；第四节是本文的重点，介绍知识抽取的算法；第五节是实验；最后是相关工作和结论。

2. 本文要解决的问题

2.1. 两个引例

本小节通过两个简单的案例来直观地阐述本文要解决的问题。假设有文本片段：“周恩来总理出生于1898 年 3 月 5 日，逝世于 1976 年 1 月 8 日”。使用命名实体识别技术可以将“1898 年 3 月 5 日”和“1976 年 1 月 8 日”标识为一个“日期时间”实体类型的实例，但不能确定“1898 年 3 月 5 日”和“1976 年 1 月 8 日”是“出生时间”还是“逝世时间”。使用实体关系抽取技术可以确定“(周恩来，1898 年 3 月 5 日)”是“出生(人物，时间)”关系的实例，“(周恩来，1976 年 1 月 8 日)”是“逝世(人物，时间)”关系的实例。通过确定实体的关系，“1898 年 3 月 5 日”和“1976 年 1 月 8 日”两个实体就有了更为明确的语义。

但在某些情况下，二元实体关系抽取的结果也不能保证实体的语义完全明确。假设有文本片段：“英华公司 2008 年实现销售收入 20.1 亿元，而 2009 年则达到 22.3 亿元”。使用命名实体识别技术可以将“英华公司”、“2008 年”、“20.1 亿元”、“2009 年”和“22.3 亿元”分别标识为“组织”、“日期”、“销售收入值”、“日期”和“销售收入值”实体类型的实例；进一步，使用实体关系抽取技术可以确定“(英华公司，20.1 亿元)”和“(英华公司，22.3 亿元)”是“销售收入(组织，销售收入值)”的两个关系实例。显然，这两个关系实例的语义由于缺乏日期实体支持而变得模糊。

笔者认为，使得一个实体具有明确语义的关键在于能够确定该实体在领域中的“位置”。如果使用领域本体来描述一个领域里的规则，那么知识抽取的问题就转换为：在领域本体中找到一个最合适的概念作为实体的类型。

2.2. 相关概念

首先我们给出同问题描述相关的几个概念。

定义 1：**本体** O 是一个五元组，$O = \{C, R, H^C, Rel, A^o\}$[1]，其中 C 是概念的集合；R 是非分类关系的集合；$H^C \subseteq C \times C$ 是分类关系的集合；$Rel: R \rightarrow C \times C$ 是一个函数，表示两个概念之间的特定关系；A^o 是公理集，通常使用逻辑语言来表示。表示本体的语言有多种，本文中提到的领域本体是使用 OWL DL 表示的本体。

为了描述问题的方便，我们引入约定 1。

约定 1：使用 $O.C$、$O.R$、$O.H^C$、$O.Rel$ 和 $O.A^o$ 来表示本体 O 的相应元组；$\forall c \in O.C$ 使用 $instconc(c)$ 表示 c 的所有实例，即实体；$\forall r \in O.Rel$ 使用 $instrela(r)$ 表示 r 的所有实例，即实体对。

基于领域本体约束的知识抽取方法是在命名实体识别和实体关系抽取的基础上完成的，因此我们首先给出命名实体识别和实体关系抽取的定义及其相关约定。

定义 2：**命名实体识别**主要是要识别出文本中出现的专有名词和有意义的数量短语的实体类型。命名实体识别中的实体类型同领域本体中的概念在语义上是一致的。

一般来说，在进行命名实体识别时往往会使用一套实体识别规则。为了描述问题的方便，我们引入约定 2。

约定 2：对于一个给定的领域本体 O，$\forall c \in O.C$ 都对应一个抽取规则集 $c.PAT$，$\forall I \in instconc(c)$ 使得 $\exists p \in c.PAT$，概念实例 i 满足抽取模式 p。称 i 是 c 的实例，记为 $instofconc(i, c)$。

定义 3：**实体关系抽取**是指从文本语料中自动识别出具有语义关系的实体对。实体关系抽取中的实体关系同领域本体中的 Rel 元组在语义是一致的。

一般来说，在进行实体关系抽取时往往会对应着一套抽取规则。为了描述问题的方便，我们引入约定 3。

约定 3：对于一个给定的领域本体 O，$\forall r \in O.Rel$ 都对应一个抽取规则集 $r.PAT$，$\forall I \in instrela(r)$ 使得 $\exists p \in r.PAT$ 关系实例 i 满足抽取模式 p。称 i 是 r 的实例，记为 $instofrela(i, r)$。

2.3. 问题的一般形式

图 1 所示的是一个非正式的领域本体 O 的示意

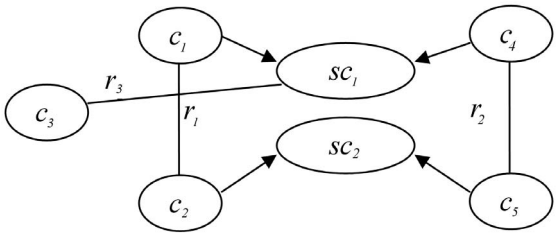

Figure 1. A diagram about ontology *O*
图 1. 领域本体 *O* 示意图

图，已知：① $c_1, c_2, c_3, c_4, c_5, c_6 \in O.C$;

② $r_1(c_1, c_2), r_2(c_4, c_5), r_3(c_1, c_3), r_4(c_5, c_6) \in O.Rel, r_1(c_1, c_2)$

表示 c_1 和 c_2 存在语义关系 $r_1, r_2(c_4, c_5)$、$r_3(c_1, c_3)$和 $r_4(c_5, c_6)$的含义类似;

③ $(c_1, sc_1), (c_4, sc_1), (c_2, sc_2), (c_5, sc_2) \in O.H^C$, (c_1, sc_1)表示 c_1 是 sc_1 的子概念，(c_4, sc_1)、(c_2, sc_2)和(c_5, sc_2)的含义类似;

④ $instconc(c_1) \cap instconc(c_4) \neq \varnothing \wedge instconc(c_2) \cap instconc(c_5) \neq \varnothing$;

⑤ $instprop(r_1) \cap instprop(r_2) \neq \varnothing$。

对于两个实体 e_1 和 e_2，如果$\exists p_1 \in c_1.PAT$，使得 e_1 满足抽取模式 p_1，那么有 $e_1 \in instconc(c_1)$成立。类似地，如果$\exists p_2 \in c_4.PAT$，使得 e_1 满足抽取模式 p_1，那么有 $e_1 \in instconc(c_4)$成立。如果$\exists p_3 \in c_2.PAT$，使得 e_2 满足抽取模式 p_3，那么有 $e_2 \in instconc(c_2)$成立。$\exists p_4 \in c_5.PAT$，使得 e_2 满足抽取模式 p_4，那么有 $e_2 \in instconc(c_5)$成立。

从上述分析可以得出结论 1) $e_1 \in instconc(sc_1)$和结论 2) $e_2 \in instconc(sc_2)$，但无法具体确定 e_1 和 e_2 所属的概念。进一步，如果$\exists p_5 \in r_1.PAT$，使得(e_1, e_2)满足抽取模式 p_5，那么有$(e_1, e_2) \in instrela(r_1)$成立，那么就可以确定 $e_1 \in instconc(c_1)$，$e_2 \in instconc(c_2)$。但如果同时又有$\exists p_6 \in r_2.PAT$，使得(e_1, e_2)满足抽取模式 p_6，那么有$(e_1, e_2) \in instrela(r_2)$成立，这样 e_1 和 e_2 所属的概念仍不明确。

但如果有实体 e_3，且已知① $e_3 \in instconc(c_3)$;

② $(e_1, e_3) \in instreal(r_3)$; ③ $O.A^o$ 中有规则

$e_1 \in instconc(sc_1) \wedge e_3 \in instconc(c_3) \wedge (e_1,e_3) \in instreal(r_3) \rightarrow e_1 \in instconc(c_1)$;

④ $O.A^o$ 中有规则

$e_1 \in instconc(c_1) \wedge ((e_1, e_2) \in instrela(r_1) \vee (e_1, e_2) \in instreal(r_2)) \rightarrow e_2 \in instconc(c_2)$;

那么进而可以确定 $e_1 \in instconc(c_1)$和 $e_2 \in instconc(c_2)$。

从图 1 我们可以得到一个启示：在经过命名实体识别和实体关系抽取之后，我们可以借助本体中已有的规则和本体提供的逻辑推理机制来进一步缩小实体所属概念的范围。下面我们将围绕这一思想给出基于领域本体约束的知识抽取算法。

3. Alloy 语言与基于本体的知识抽取方法

基于领域本体约束的知识抽取流程如图 2 所示。因为流程中用到了 Alloy 语言和 Alloy 分析器，因此我们首先对 Alloy 语言和 Alloy 分析器作简要的介绍。

3.1. Alloy 语言与 Alloy 分析器[2-4]

Alloy 语言是基于关系逻辑的结构化建模语言，用于对具有复杂结构和行为的系统进行建模。Alloy 语言用于描述的对象是原子和元组(原子序列)。下面通过一个简单的例子简要地介绍 Alloy 语言。如图 3 所示：

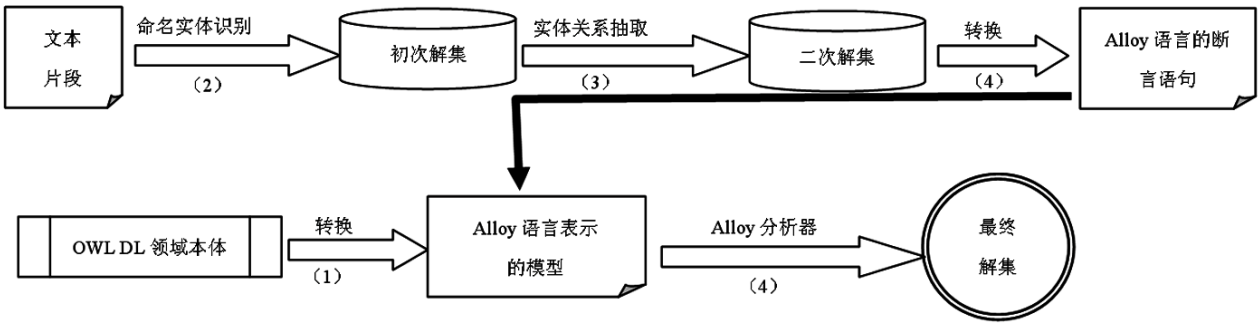

Figure 2. An approach of knowledge extraction restrained by ontology
图 2. 基于领域本体约束的知识抽取方法

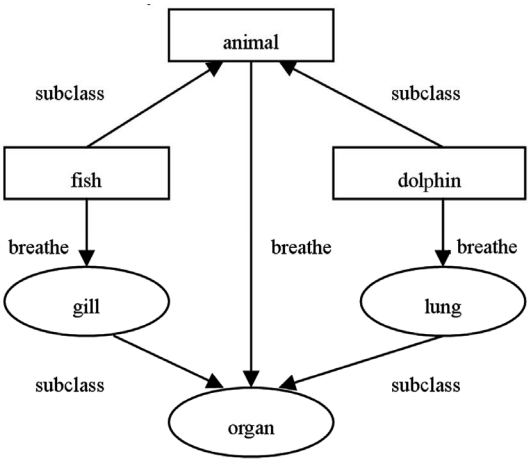

Figure 3. A case about animal
图 3. 一个描述动物呼吸的案例

fish 和 dolphin 都是 animal 的子类，fish 用 gill 呼吸，dolphin 用 lung 呼吸，gill 和 lung 都是 organ 的子类，animal 使用 organ 呼吸。可以使用 Alloy 语言(图 4 所示)来形式化地表示上述事实。在图 4 中，行 1 表示模型的名字是“animal”，“module”是 Alloy 语言中的关键字；行 2 表示“animal”集合是“Class”集合的子集，“sig”关键字用于定义一个集合；类似地，行 3 至行 5 分别定义了“gill”、“fish”和“breathe”集合，其中“gill”和“fish”是 Class 的子集，“breathe-by”是“Property”的子集；行 6 陈述事实(fact)：“fish”是“animal”的子集；行 7 陈述事实：所有的“fish”用“gill”呼吸(breathe)；行 8 表示 lung 是 Class 的子集；行 9 表示 lung 与 gill 的交集为空；行 10 表示 dolphin 是 Class 的子集；行 11 表示 dolphin 是 animal 的子集；行 12 表示所有的 dolphin 都用 lung 呼吸(breathe)；行 13 陈述一个断言 dolphinIsFish：dolphin 是 fish 的子集；行 14 根据已知的事实和前提来验证断言 dolphinIsFish 的正确性。将图 4 的代码调入到 Alloy 分析器并设定相关的参数后，Alloy 分析器输出对断言 dolphinIsFish 的验证结果。

3.2. 本体与 Alloy[5]

以描述逻辑为基础的本体描述语言 OWL DL 所表达的本体信息可以用一个 Alloy 模型来描述。一个领域本体可以映射为 Alloy 语言的一个模型，本体中的概念可以映射为 Alloy 语言中的集合；本体中关系可以映射为集合间的关系；本体中的概念层次可以映射为 Alloy 语言中的子集关系；本体中的公理则完全可以用 Alloy 语言中的 fact 来陈述。

3.3. 基于领域本体约束的知识抽取方法

第一步：将 OWL DL 语言描述的本体转换为一个用 Alloy 语言表达的模型 M^*。

第二步：借助命名实体识别技术确定原始文本片段的初次解集。

记 $E=\{e_1,e_2,\cdots,e_n\}$ 为原始文本片段中所有待识别的实体集合；$\forall e \in E$ 使用命名实体识别技术可以得到 e 所有可能的实体类型，这些类型称为 e 的实体目标域。形式化地，我们给出实体目标域和初次解集的定义。

定义 4：$\forall c \in O.C$，称

$e.c = \{c | c \in O.C, instofconc(e, c)\}$为实体 e 的实体目标域。

定义 5：称 $space(E)^1 = e_1.c \times e_2.c \times \cdots \times e_n.c$ 为 E 的初次解集。

第三步：在实体关系抽取的基础上，使用图 5 所示的算法约简初次解集得到二次解集 $space(E)^2$。

定义 6：$\forall (c_a, c_b) \in O.R$，称

$(e_1, e_2).r = \{(c_a, c_b) | (c_a, c_b) \in O.R, instofprop((e_1, e_2), (c_a, c_b))\}$

为实体对(e_1, e_2)的关系目标域。

```
1   module animal
2   sig animal extends Class {}
3   sig gill extends Class {}
4   sig fish extends Class {}
5   sig breathe extends Property {}
6   fact {subClassOf[animal, fish] }
7   fact {allValuesFrom[breathe, fish, gill] }
8   sig lung extends Class {}
9   fact {disjointWith[lung, gill] }
10  sig dolphin extends Class {}
11  fact {subClassOf[animal, dolphin] }
12  fact {allValuesFrom[breathe, dolphin, lung] }
13  assert dolphinIsFish { subClassOf[fish, dolphin] }
14  check dolphinIsFish for 6
```

Figure 4. A case for Alloy model
图 4. Alloy 模型示例


```
输入：
■  space(E)^1
输出：space(E)^2   //二次解集
方法：getSpace^2
(1) space(E)^2←space(E)^1
(2) for each (e_i,e_j) in E×E
(3)   for each (c_1,c_2,…c_i,…,c_j,…,c_n) in space(E)^1
(4)     if (c_i,c_j)∉(e_i,e_j).r and (e_i,e_j).r≠∅ then
(5)       space(E)^2←space(E)^2-{c_1,c_2,…c_i,…,c_j,…,c_n}
(6)     end if
(7)   end for
(8) end for
```

Figure 5. The algorithm for solving second answers set
图 5. 求二次解集算法

第四步：将二次解集中的每一个解转换为 Alloy 模型下一组断言语句，然后将这组断言语句补充到 $M^{\times}$；如果 Alloy 分析器输出每一个断言都是有效的，那么认为这组断言对应的解是一组有效解；否则是无效解。算法如图 6 所示。

4. 实验

4.1. 实验数据

为了验证算法的有效性，我们以人物讣告主题的文本片段作为实验对象。首先我们定义了如下的本体 O，其中

- C = {*Person*, *Deceased Person*, *Relative*, *Age*, *Birth Date*, *Death Date*, *Death Address*, *Birth Address*, *Date*, *Address*}，*Person* 代表“人”，*Date* 代表“日期”，*Address* 代表“地点”；*Deceased Person* 代表“死者”，*Relative* 代表“亲属”，*Age* 代表“卒年”，*Birth Date* 表示出生日期，*Death Date* 代表死亡日期，*Death Address* 代表死亡地点，*Birth Address* 代表出生地点；
- $O.R$ = {*born in address*, *die at the age of*, *born in date*, *die in date*, *die in address*, *is related to*}；
- $O.H^C$ = {(*Deceased Person*, *Person*), (*Relative*, *Person*), (*Birth Address*, *Address*), (*Death Address*, *Address*), (*Birth Date*, *Date*), (*Death Date*, *Date*)}；
- $O.Rel$ = {*born in address* (*Deceased Person*, *Birth Address*), *die at the age of* (*Deceased Person*, *Age*), *born in date* (*Deceased Person*, *Birth Date*), *die in date* (*Deceased Person*, *Death Date*), *die in address* (*Deceased Person*, *Death Address*), *is related to* (*Deceased Person*, *Relative*)}；
- $O.A^o$ = {$e_1 \in$ *instconc* (*Deceased Person*) $\wedge$ $e_2 \in$ *instconc*(*Date*) $\wedge$ (e_1, e_2) $\in$ *instreal* (*born in date*) $\rightarrow e_2 \in$ *instconc* (*Birth Date*), $e_1 \in$ *instconc* (*Deceased Person*) $\wedge$ $e_2 \in$ *instconc* (*Date*) $\wedge$ (e_1, e_2) $\in$ *instreal* (*die in date*) $\rightarrow e_2 \in$ *instconc*(*Death Date*), $e_1 \in$ *instconc* (*Deceased Person*) $\wedge$ $e_2 \in$ *instconc* (*Address*) $\wedge$ (e_1, e_2) $\in$ *instreal* (*born in Address*) $\rightarrow e_2 \in$ *instconc* (*Birth Address*), $e_1 \in$ *instconc* (*Deceased Person*) $\wedge$ $e_2 \in$ *instconc* (*Address*) $\wedge$(e_1, e_2) $\in$ *instreal* (*die in Address*) $\rightarrow e_2 \in$ *instconc* (*die Address*), $e \in$ *instconc* (*Deceased Person*) $\rightarrow e_2 \notin$ *instconc* (*Relative*), $e_1 \in$ *instconc* (*Deceased Person*) $\wedge$ $e_2 \in$ *instconc* (*Deceased Person*) $\rightarrow e_1 = e_2$}

然后我们从 Salt Lake Tribune(www.sltrib.com)和 Arizona Daily Star(www.azstarnet.com)下载了 30 篇讣告进行测试。

4.2. 实验结果

从前三节的描述里不难发现，最终解集中解的数量越少，说明实验结果越精确；反之则实验的不确定性越高。理想化地，对于每次抽取结果，最终解集中解的数量应为 1。而从实验结果(如图 7 所示)来看，只有 6 篇讣告最终解集中解的数量大于 1，其余 24 篇都获得了明确抽取结果。因此，实验结果验证了新方法的有效性。

```
输入：
■  M*   //Alloy 模型
■  space(E)^2
输出：space(E)^2
方法：getResult
(1)for each s=(c_1,c_2,…,c_n) in space(E)^2
(2)    i←1;
(3)    while(i<=n)
(4)      M←M*⊙"assert isInstance{e_i in c_i.instances}";
(5)      //为模型 M*补充断言语句
(6)      if Alloy 分析器认为模型 M 的断言为假 then
(7)        space(E)^2←space(E)^2-{s}
(8)        break;
(9)      else
(10)       i←i+1;
(11)     end if
(12)   end while
(13)end for
```

Figure 6. The algorithm for solving final answers set
图 6. 求最终解集算法

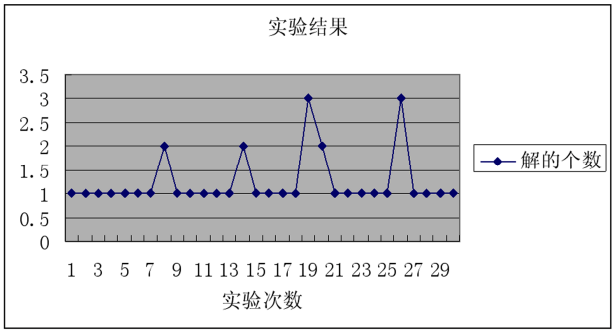

Figure 7. Experiment result
图 7. 实验结果

5. 相关工作

[6]提出一种由 Record Extractor、Constanst/Keyword Recognizer、Ontology Parser、Database-Instance Generator 四大部件构成的信息抽取系统。系统的工作原理是首先使用 Record Extractor 从原始网页中提取含有待抽取信息项的记录条；然后使用 Ontology Parser 从领域本体中获取信息项的特征知识；根据信息项的特征知识和Constanst/Keyword Recognizer中包含的约束信息从记录条中抽取信息项，最后依使用 Database-Instance Generator 将信息项存到数据库中。系统的关键部件是 Constanst/Keyword Recognizer 和 Ontology Parser，它们提供了信息抽取的规则，这些规则都是基于启发式或者预定义的语法格式抽取规则，因此该方法缺乏通用性。此后，[7,8]从不同的应用背景出发开发了各自的基于本体的信息抽取系统，但其指导思想同[6]相比基本一致。[7]以商务智能领域作为背景，在领域本体的支持下，使用基于语言学模式的的方法进行信息抽取。Kylin system[8]以 Wikipedia pages 中包含的信息作为抽取对象，利用 WordNet 中提供的概念以及概念间的关系来构造本体，使用最大熵模型和 CRF 来对语句中包含的信息进行识别。由于该系统使用统计的方法来进行信息抽取，因此对于缺乏足够训练样例的情况抽取效果较差。

6. 结束语

本文提出了一种基于本体约束的知识抽取的方法。不同于传统的基于规则的信息抽取方法，我们从知识的角度出发，将知识抽取转换为一个约束逻辑程序求解问题，不仅合成了已有的知识片段，获得了完整的知识实例，而且利用本体中定义的逻辑规则，克服了不同的信息项满足同一或具有包含关系抽取规则时无法确定信息项类型的问题。

参考文献 (References)

[1] L. Stojanovic. Methods and tools for ontology evolution. University of Karlsruhe, 2004.
[2] D. Jackson, I. Schechter and I. Shlyakhter. ALCOA: The alloy constraint analyzer. Proceeding of 22nd International Conference on Software Engineering (ICSE), 2000. ACM Press, 2000: 331-347.
[3] D. Jackson. Micromodels of software: Modelling and analysis with Alloy. http://sdg.lcs.mit.edu/alloy/book.pdf.
[4] D. Jackson. Alloy: A lightweight object modeling notation. ACM Transactions on Software Engineering and Methodology (TOSEM), 2002, 11(2): 256-290.
[5] H. H. Wang. Reasoning support for Semantic Web ontology family languages using alloy. Multiagent and Grid Systems, 2006, 2(4): 145-155.
[6] D. W. Embley. Conceptual-model-based data extraction from multiple-record web pages. Data & Knowledge Engineering, 1998, 31: 227-251.
[7] H. Saggion, A. Funk, D. Maynard and K. Bontcheva. Ontology-based information extraction for business intelligence. Proceedings of the 6th International the Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference, 2007, 6(3): 843-856.
[8] F. Wu, R. Hoffmann and D. S. Weld. Information extraction from Wikipedia: Moving down the long tail. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, 5(1): 731-739.

