

# 基于多特征表示的本体概念挂载\*

徐立恒<sup>1</sup>, 刘洋<sup>1</sup>, 来斯惟<sup>1</sup>, 刘康<sup>1</sup>, 田野<sup>2</sup>, 王渝丽<sup>2</sup>, 赵军<sup>1</sup>

<sup>1</sup>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

<sup>2</sup>中国大百科全书出版社, 北京 100037

E-mail: {lhxu, liuyang09, swlai, kliu, jzhao}@nlpr.ia.ac.cn

**摘要:** 本文研究了一种基于多特征表示的本体概念挂载方法。以中国大百科知识体系\*\*作为本体体系结构, 抽取网络知识库条目作为本体概念, 通过分析条目标文内容、语义标签和半结构化信息获得概念间层级关系。本文将中国大百科知识体系扩展为百万级概念的多领域中文本体, 为进一步抽取本体概念属性、概念之间的非层级关系及问答服务等应用建立良好的基础。实验证明本文方法相对于单一特征方法能够提高 11.8% 的挂载精度。

**关键词:** 本体; 多特征; 概念挂载

## An Ontology Concept Acquisition Method Based on Heterogeneous Features

Xu Liheng<sup>1</sup>, Liu Yang<sup>1</sup>, Lai Siwei<sup>1</sup>, Liu Kang<sup>1</sup>, Tian Ye<sup>2</sup>, Wang Yuli<sup>2</sup>, Zhao Jun<sup>1</sup>

<sup>1</sup>National Laboratory of Pattern Recognition, Institute of Automation, The Chinese Academy of Sciences, Beijing 100190

<sup>2</sup>Encyclopedia of China Publishing House, Beijing 100037

E-mail: {lhxu, liuyang09, swlai, kliu, jzhao}@nlpr.ia.ac.cn

**Abstract:** This Paper proposes an ontology concept acquisition method based on heterogeneous features. We regard the Encyclopedia of China as the taxonomy of ontology, extract Web knowledge base articles as concepts and learn taxonomic relations between concepts by considering the text content, folksonomies as well as semi-structured information. We extend the Encyclopedia of China to a mega-scale global Chinese ontology which provide practical support for concept attributes extraction, non-taxonomic relations extraction and other applications such as Question Answering System. Experimental results show that the proposed method achieved 11.8% performance improvement compared to the single feature method.

**Keywords:** ontology; heterogeneous features; concept acquisition

### 1 引言

近年来, 本体被广泛应用于信息集成、智能信息检索、自然语言处理等领域, 并被视为语义网应用和解决异构信息系统互操作问题的关键技术之一。现有本体如 CYC<sup>[1]</sup>和 HowNet<sup>[2]</sup>等大多依靠专家知识人工编撰。随着知识呈爆炸式地增长, 本体创建已经遇到了知识获取瓶颈: 手工编撰不仅费时费力, 而且知识覆盖率低, 数据稀疏, 更新缓慢。因此, 有必要发展监督或半监督的自动本体构建方法。人工协同创建本体是现阶段较可行的方法之一, 同时 Wikipedia、Freebase<sup>[3]</sup>等大规模知识库的迅速发展, 为基于网络知识库的本体构建新方法<sup>[4][5]</sup>提供了契机。

目前中文网络知识库(如维基百科中文版)没有一个规范统一的分类体系, 无法被用作实际应用系统的支撑。中国大百科提供了一个由专家制定的权威的体系结构, 但该知识体系概念数量较少。本文使用中国大百科知识体系作为目标本体的分类体系, 从网络知识库抽取概念, 并将概念实例挂载到大百科知识体系的层级结构中, 从而构建一个分类体系规范的海量中文本体。传统方法往往认为这个过程是一个文本分类问题, 而忽略了网络百科条目中所包含的半结构化信息与

\* 本文受国家自然科学基金项目(60875041, 61070106)和中国出版集团科技项目资助。

\*\* 中国大百科知识体系是目前定义最完善、知识覆盖面最广的中文百科知识体系; 是一个拥有 55 个一级类、78292 个知识条目的树结构。

语义信息。因此本文提出了一种多特征表示的本体概念挂载方法。我们的主要贡献在于：

1. 提出了一种融合概念的文本内容、语义标签和半结构化特征判断概念类别的方法。通过网络知识库条目中多特征的综合描述，能够有效地捕捉条目之间语义关联。
2. 构建了一个百万规模概念的多领域中文本体，并为下一步抽取本体概念属性、概念之间的非层级关系及问答服务等应用建立了良好的基础。

本文的其他部分按照如下安排：第二节简要地回顾本体构建与本体概念挂载相关工作；第三节详细描述了我们的工作及方法；第四节展示我们构建的大规模中文本体的相关性能与实验结果；第五节中我们对已完成的工作进行综述并对将来工作进行了展望。

## 2 相关工作

目前本体构建方法主要有如下三种：基于结构化数据的方法、基于非结构化或半结构化数据的方法<sup>[13]</sup>和基于网络知识库的方法。

基于结构化数据的本体构建主要从关系数据库或面向对象数据库中获取概念。这类方法主要通过关系模式进行语义分析，利用规则获取本体概念和概念间关系，如 Man Li 等<sup>[6]</sup>的工作。目前基于结构化数据的本体构建大多使用规则方法，因此扩展性较差。

基于非结构化或半结构化数据的本体构建主要从纯文本或半结构化网页中获取概念。主要方法有：模板方法、概念聚类方法和机器学习方法。模板方法的经典模式是使用 Hearst 模板<sup>[7]</sup>（形如“NP such as {NP,}\* (and/or) NP”等）从纯文本语料中获取概念实例。这类方法的主要缺点是获取候选概念的准确度低，因为大量无用的概念往往也会匹配这些模式。概念聚类方法是利用同一类概念之间的语义近似关系对概念进行聚类，如 Bisson 和 Emde 等人<sup>[8]</sup>等提出的基于 FOL 的聚类方法。聚类方法对概念间的层级关系抽取效果较好，但是候选概念却需要人工输入或使用其他工具获得。机器学习方法通常利用监督或半监督方法学习分类模型并获取概念。Cimianonce 等人<sup>[9]</sup>融合了模板、句法结构和 WordNet 层级分类结构等多种信息，使用多种分类器比较了获取概念并生成层级结构的效果，但是其分类效果最好的 SVM 分类器取得的 F 值仅 33%。

随着网络知识库日渐成熟，基于网络知识库的本体构建方法成为主流技术。DBpedia<sup>[5]</sup>通过分析 Wikipedia 网页中的 Category Page、Infobox 等半结构化信息将概念归入人工定义的知识体系中，获得约 260 万概念。YAGO<sup>[4]</sup>以 WordNet 的层级结构作为知识体系，使用启发式规则从 Wikipedia 中抽取概念并挂载到 WordNet 类（synset）中，获得了超过 100 万个概念。这两项工作为我们提供了一种良好的本体构建框架：即用一种人工定义的、体系结构良好的小规模本体作为核心，向其中大规模添加概念并获得概念间关系。下文中我们采用这种框架进行大规模中文本体概念挂载。

## 3 基于多特征表示的概念挂载

本体概念挂载包含知识体系构建、候选概念抽取和概念间层级关系抽取三部分。本文的方法中，知识体系已由专家构建完成，候选概念抽取自网络百科条目，我们将工作的重点放在概念间层级关系的抽取上，即新增概念实例与大百科知识体系概念之间的层级关系抽取。我们综合利用了网络百科条目的多种分类特征信息抽取概念之间的层级关系。

### 3.1 动机

图 1 是一个网络百科条目示例，图中可以表达该条目类别的信息有如下几部分：

1. 标题：如标题“鲁迅”说明该条目可能是一个人。
2. 正文：正文内容详细介绍了该知识条目。
3. 属性框：属性框（Infobox）是描述条目属性信息的表格。图中右侧属性框中的“出生”、“配

偶”等属性名强烈提示该条目是一个人。

4. 开放分类: 开放分类 (Folksonomies) 是用户使用开放标签对条目的归类信息。图中底部开放分类文本直接指明了“鲁迅”属于“中国现代作家”。
5. 相关词条: 条目中大量 HTML 链接指示了与“鲁迅”相关的其他条目。

The image shows a screenshot of a Wikipedia article for Lu Xun (鲁迅). The article is in Chinese and includes a title bar, a main text area, a sidebar with a portrait, and a bottom section for open classifications. The main text area contains a brief biography and a section titled '鲁迅的朋友和学生'. The sidebar includes a portrait of Lu Xun and a table of personal information. The bottom section lists open classifications such as '中国现代作家' and '中国小说家'.

图1 网络百科知识条目示例

在分析了大量类似的百科知识条目网页后, 针对概念间层级关系抽取问题, 我们的思路大致如下:

1. 利用概念之间的语义相似度推断概念所属的类别。如描述“鲁迅”的概念通常会与“郭沫若”、“郁达夫”等概念在正文、开放分类、相关词条等内容上相似。表1是使用3.3节中概念语义相似度计算方法得到的不同类别概念之间的相似度示例。
2. 利用上文中所述的多种特征信息综合推断概念所属的类别。不同的特征表达概念类别的信息量不同, 如属性框信息中属性“职业”的属性值是提示概念分类的直接证据, 而开放分类标签数量多、歧义大, 很难分析出主要分类标签。因此我们使用统计分类算法结合启发式规则的方法适应不同强度的特征。分类算法利用概念之间的语义相似度对概念归类, 并限制规则适用范围; 启发式规则用于发现提示概念类别的直接证据, 并监督指导分类算法。

表1 概念间相似度得分示例

相似度得分	郭沫若	郁达夫	绍兴	黄酒
鲁迅	0.543	0.599	0.330	0.330
乐山	0.352	0.328	0.560	0.332
啤酒	0.297	0.310	0.299	0.524

根据以上思路, 本文使用层级kNN算法结合启发式规则方法实现本体概念间层级关系抽取。下文将对本体概念挂载方法进行详细介绍。

### 3.2 层级分类算法

我们将大百科知识库层级体系定义为  $T$ , 所有概念实例集合定义为  $S$ , 需要被添加进入大百科体系的网络知识条目网页集合定义为  $D$ 。算法输入待分类条目  $d \in D$ , 输出一个大百科类  $c \in T$  且知识条目  $d$  是类  $c$  的一个实例。算法核心思想是将条目  $d$  在树结构  $T$  中自上而下逐层分类, 在每一层使用一次kNN分类与启发式规则相结合的方法。图2是层级分类算法伪代码。

---

```

Function get_category(d)
Input: an article d
Output: the category of d
1:   c ← root of tree T
2:   calculate each semantic similar score w(d, s) in which s ∈ S
3:   while c has subclasses
4:     S' ← all children of c in S
5:     Q ← a queue consist of w(d, s) in which s ∈ S' and w(d, s) > θ
6:     process kNN algorithm with heuristic rules on Q to get a class c' ∈ T and d is an instance of c'
7:     c ← c'
8:   return c

```

---

图2 层级分类算法伪代码

### 3.3 概念语义相似度计算

概念之间的语义相似度计算是分类算法的关键。由于大百科知识体系中原有的概念没有语义信息与半结构化信息，首先需要使用函数  $f(s)$  将大百科概念  $s$  通过简单名称匹配映射为网络知识条目。此后，我们融合了知识条目的正文信息 (BW)、开放分类信息 (TG) 和相关词条信息 (RL) 计算语义相似度，计算方法如式(1)所示，其中  $\alpha$ 、 $\beta$ 、 $\gamma$  为加权值。

$$w(d, s) = \alpha * BW(d, s) + \beta * TG(d, s) + \gamma * RL(d, s) \quad (1)$$

下面介绍语义相似度值各部分的计算方法。

#### 3.3.1 正文相似度计算

将网络知识条目的正文所有内容看作一个词袋，则可用向量  $v(t_1, t_2, \dots, t_n)$  表示正文，其中  $t_i$  是词  $i$  在条目正文中的 *tf-idf* 值。我们使用工具<sup>[10]</sup>分词，在进行相似度计算时，只保留名词和动词。则条目  $d$  与概念  $s$  映射后的条目  $f(s)$  的正文相似度为：

$$BW(d, s) = \cos(v_d, v_{f(s)}) \quad (2)$$

#### 3.3.2 开放分类相似度计算

开放分类指用户对知识条目的归类标签。由于标签文本是开放的，因此同一个类的条目会拥有不同的标签文本。如描述地理类条目的标签可能有“地理”、“地点”、“地名”等多种形式，简单的字符串匹配无法处理意思相同而形式不同的标签。L. Specia 等在<sup>[11]</sup>中提供了一种计算标签之间两两相似度的方法：设所有标签集合为  $T$ ，构建一个  $T \times T$  的共现矩阵  $P$ ，其中  $P_{ij}$  为标签  $t_i$  与  $t_j$  在条目中的共现次数。则标签  $t_m$  与  $t_n$  的相似度得分为向量  $P_m$  与  $P_n$  的  $\cos$  值，从而可以得到一个  $T \times T$  的标签相似度矩阵  $M$ ，其中  $M_{mn}$  为 1.0。设条目  $d$  与概念  $s$  映射后的条目  $f(s)$  的开放分类标签集合分别为  $T_d$  与  $T_{f(s)}$ ，两者交集为  $U$ ， $T_d$  与  $U$  的差集为  $I_d$ ， $T_{f(s)}$  与  $U$  的差集为  $I_{f(s)}$ ，则

$$TG(d, s) = \frac{\sum_{t_m \in I_d, t_n \in I_{f(s)}} (M_{mn}) + (|T_d| * |T_{f(s)}| - |I_d| * |I_{f(s)}|)}{|T_d| * |T_{f(s)}|} \quad (3)$$

#### 3.3.3 相关词条相似度计算

相关词条是知识条目网页描述中用户标出的与之相关的同类知识条目，同时条目正文中的相互链接也可以视为相关词条。设  $r_{f(s)}(d_1, d_2, \dots, d_n)$  表示概念  $s$  的相关词条向量，其中  $d_1, d_2, \dots, d_n$  是集合  $D$  中的所有条目。条目  $d_i$  与  $s$  相关定义为概念  $s$  映射为条目  $f(s)$  后的正文中存在指向  $d_i$  的 HTML 链接。则  $r_{f(s)}$  中与  $s$  相关的条目  $d_i$  的值为 1，与  $s$  不相关的条目  $d_j$  的值为 0，则

$$RL(d, s) = \cos(r_d, r_{f(s)}) \quad (4)$$

### 3.4 启发式规则

统计分类方法容易湮没知识条目中表征条目类别的直接证据，因此我们在统计方法的基础上还利用了一些启发式规则来识别表示条目类别信息的直接证据。同时，启发式规则的应用范围由统计分类算法限制，以提高规则的准确率。下面按照优先级从高到低顺序详细介绍启发式规则。

#### 3.4.1 属性框规则

当检测到待分类条目  $d$  包含属性框时，分类规则自动抽取属性框中的属性名与对应属性值。属性框规则分为两类：1) 粗分类规则：由多个属性名及一个大百科第一级类组成。若一个条目的属性框中抽取的属性名与粗分类规则的属性名相符个数大于等于三个，则判定该条目属于规则指定类。2) 细分类规则：对应一条粗分类规则并由多个属性名构成。在判断一个条目满足某条粗分类规则后，检测条目的属性框中是否包含对应的细分类规则的属性名。若包含，则直接使用属性名对应的属性值与统计分类结果的前三个类名进行匹配，判断条目所属的具体类别；若不包含或属性值无法匹配类名，则仍使用层级分类结果。表 2 中是一些属性框规则示例。

表 2 属性框规则示例

第一级类	属性名个数	粗分类规则属性名	细分类规则属性名
人物	32	国籍, 民族, 身高, 体重, 出生……	职业, 运动项目, 转入职业……
地理	46	面积, 人口, 电话区号, 时区……	地理位置, 所属地区……
教育	17	学校类型, 现任校长, 校色, 校址……	类别, 学校类型, 学校属性……
生物	7	界, 门, 纲, 目, 科, 属, 种	界, 门, 纲, 目, 科, 属, 种

#### 3.4.2 命名实体规则

初步统计网络知识条目的类别，我们发现命名实体（人名、地名、机构名）占条目总数约 20% 以上。如果可以较准确识别命名实体，那么可以有效阻止机构被识别为 人物、人物被识别为著作等情况。我们使用工具<sup>[10]</sup>识别命名实体。若一个条目的标题在正文中被识别为命名实体的比率超过 75%，则认为该条目是一个命名实体。在分类时，人工指定某些类为命名实体类，在判断第一级类时仍使用分类算法，在下层分类中根据规则将被识别为命名实体的条目分入相应的命名实体类中。表 3 中是一些命名实体规则示例。

表 3 命名实体规则示例

命名实体类别	适用类数	适用类名
PERSON	52	文学->作家, 医学->医生, 物理->物理学家 ……
LOCATION	243	地理->国家, 地理->地名, 地理->名胜古迹, 考古->遗址 ……
ORGANIZATION	94	物理->组织机构, 戏剧->剧团, 医学->医院 ……

#### 3.4.3 定义句规则

正文中用于定义知识条目是什么的句子被称作定义句。定义句中往往会直接给出知识条目的分类信息。我们使用句法分析工具<sup>[12]</sup>对正文首段中所有包含条目标题的句子进行句法分析，若句子满足定义句规则并且规则抽取得到的分类结果存在于层级分类算法每一层的前三个结果中，则分类器采用定义句规则抽取的结果。表 4 中是一些定义句规则示例。

表 4 定义句规则示例

满足的类	中心词	定义句规则
ALL	是, 属于, ……	NP([title]) [中心词] NP([category])
ALL	称为, 称作, 叫做……	NP([category]) [中心词] NP([title])
地理	位于, 坐落, ……	NP([title]) [中心词] NP([LOCATION])

## 4 实验与结果

### 4.1 实验数据与设置

我们从网络百科知识库爬取知识条目 147 万条构成上文中所述的待挂载条目集合  $D$ 。大百科知识库经过人工整理得到一级类 45 个，类 5263 个，概念实例 78292 个，选取实例数目较多的大类共 851 个构成层级体系  $T$ 。大百科知识库中有 58032 个概念通过  $f(s)$  函数映射到网络百科知识条目，这些概念构成语义相似度计算概念实例集合  $S$ 。我们设置了两组对比实验：1) 对比正文 (BW)、开放分类 (TG) 和相关词条 (RL) 三个特征单独使用与三个特征融合的挂载效果；2) 对比只用统计分类算法的挂载效果与统计分类算法和启发式规则结合的挂载效果。

实验参数设置如下：阈值根据经验指定为  $\theta=0.35$ ， $k$ 、 $\alpha$ 、 $\beta$ 、 $\gamma$  值由概念集  $S$  上的集内  $kNN$  分类封闭测试得到，取  $k=17$ ，令  $\alpha+\beta+\gamma=1$ ，则  $\alpha=0.57$ ， $\beta=0.12$ ， $\gamma=0.31$ 。

### 4.2 实验结果与分析

对比实验一在概念集  $S$  上进行层级  $kNN$  分类封闭测试。表 5 显示使用不同特征的每一层分类的准确率。由表可知三种特征融合后比单独使用正文文本信息的准确率提升了 11.8 个百分点。

表 5 语义相似度衡量方法效果对比

准确率 (%)	BW	TG	RL	BW+TG+RL	$\alpha*BW+\beta*TG+\gamma*RL$
第一层	82.9	68.3	79.3	87.6	91.8
第二层	76.3	63.8	74.7	81.0	86.7
第三层	71.5	59.4	70.3	77.8	82.9
最底层	67.8	58.3	64.7	75.1	79.6

对比实验二是大规模挂载网络百科条目的开放测试，共挂载条目 974984 个。我们采用随机抽取结果并人工标注的方式评价结果。对大百科每个大类按挂载比例随机抽取共 4347 个新增概念进行标注。结果如表 6 所示，其中 BS 为只使用语义相似度统计分类结果，HR 为结合规则方法所得结果。由表可知分类算法结合启发式规则方法比单纯使用分类方法准确率提高了 7.6%。

表 6 概念挂载准确率

准确率 (%)	地理		军事		生物		医学		农业		化学		全部	
	BS	HR	BS	HR	BS	HR	BS	HR	BS	HR	BS	HR	BS	HR
第一层	90.4	<b>91.0</b>	92.2	<b>92.2</b>	91.4	<b>92.1</b>	93.5	<b>94.8</b>	89.5	<b>90.0</b>	85.1	<b>85.3</b>	88.3	<b>89.4</b>
第二层	81.9	<b>84.8</b>	80.9	<b>84.8</b>	82.0	<b>83.1</b>	79.0	<b>83.4</b>	84.3	<b>86.7</b>	80.0	<b>82.1</b>	76.7	<b>79.6</b>
第三层	70.7	<b>79.3</b>	76.7	<b>81.6</b>	75.2	<b>77.4</b>	77.6	<b>82.3</b>	82.9	<b>85.3</b>	73.0	<b>75.1</b>	70.5	<b>75.0</b>
最底层	59.0	<b>75.4</b>	76.7	<b>81.6</b>	59.8	<b>73.3</b>	77.6	<b>82.3</b>	80.9	<b>83.3</b>	66.8	<b>68.9</b>	64.7	<b>72.3</b>

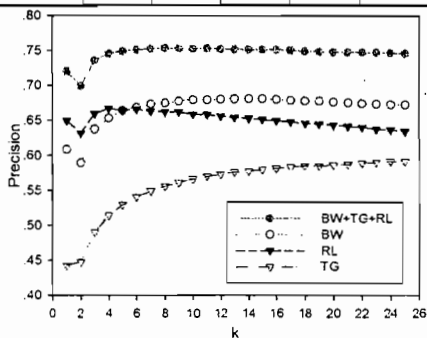


图 3 不同  $k$  值下准确率

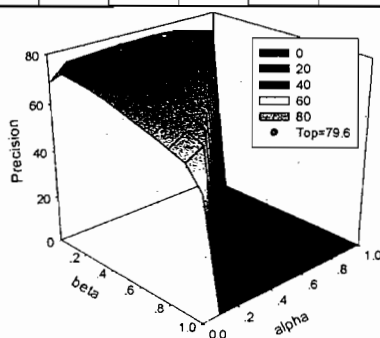


图 4 不同权重下准确率

我们对比了实验参数设置对概念挂载准确率的影响。图 3 是不同 k 值下对 S 集封闭分类测试的准确率。图 4 是的使用不同权重参数下的准确率对比结果。由图与表 5 可知采用合适加权重后的分类准确率可以提升 4.5%。

对概念挂载错误的实例进行分析可知错误原因主要有如下两类：1) 相似的类之间难以区分导致分类错误。如“医学->药品”与“化学”、“生物->植物界”与“农业->蔬菜”等。2) 分类体系数据稀疏导致分类错误。如我们的分类体系中“交通->船舶”只有 5 个概念实例，而“军事->军舰”有 56 个实例，则使用分类算法时“船舶”容易被分类为“军舰”。

## 5 结论与展望

本文研究了一种多特征表示的本地概念挂载技术，并成功将中国大百科分类体系扩展为一个拥有百万级概念的全领域中文本体。在本体概念挂载时，限定了一个概念只能拥有一个父节点，未考虑实体消歧，同时也未使用共指消解技术，这两项工作是我们后续改进工作的重点。另外，本体构建不仅需要概念挂载技术，也需要抽取概念的属性以及概念之间的非层级关系。下一步工作可以在本文基础上进行概念属性与关系的抽取。

## 参考文献

- [1] D. B. Lenat, R. V. Guha, K. Pittman, D. Pratt, and M. Shepherd. CYC: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM* 38(11): 33-38, Nov. 1995.
- [2] Dong, Z. and Dong, Q. 2000. HowNet [online]. Available at <http://www.keenage.com/>
- [3] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In SIGMOD '08, 1247-1250, New York, NY, ACM.
- [4] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A Core of Semantic Knowledge. In 16<sup>th</sup> international World Wide Web conference (WWW2007), New York, NY, USA, 2007. ACM Press.
- [5] 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.
- [6] Li, M., Du, X., Wang, S.: Learning Ontology from Relational Database. In: Proceedings of the 4<sup>th</sup> International Conference on Machine Learning and Cybernetics. Vol. 6 (2005) 3410-3415.
- [7] M. Hearst. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14<sup>th</sup> International Conference on Computational Linguistics, pages 539-545, 1992.
- [8] Bisson, G. (1992). Learning in Fol with a Similarity Measure, Tenth National Conference on Artificial Intelligence (AAAI'92), pp 82-87, San Jose, California: AAAI Press.
- [9] Cimiano, P., A. Pivk, L. Schmidt-Thieme & S. Staab (2005). Learning taxonomic relations from heterogenous sources of evidence. In P. Buitelaar, P. Cimiano & B. Magnini (Eds.), *Ontology Learning from Text: Methods, Evaluation and Applications*, pp. 59-73. Amsterdam, The Netherlands: IOS Press.
- [10] Youzheng Wu, Jun Zhao, Xu Bo, Chinese Named Entity Recognition Model Based on Multiple Features. In: Proceedings of the Joint Conference of Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, 2005: 427-434.
- [11] L. Specia and E. Motta. Integrating Folksonomies with the Semantic Web, Proceedings of the 4<sup>th</sup> European Semantic Web Conference (ESWC 2007), 2007.
- [12] Xiangyu Duan, Jun Zhao, Probabilistic Parsing Action Models for Multi-Lingual Dependency Parsing, In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-2007), Prague, Czech Republic, 940-946, 2007.
- [13] 杜小勇, 李曼, 王珊. 本体学习研究综述. *软件学报*, 2006, 17(9): 1837-1847.