第 48 卷 第 4 期
2012 年 7 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCES)

Vol. 48, No. 4
July, 2012

粗糙集与粒计算进展

# 领域本体概念实例、属性和属性值的抽取及关系预测*

郭剑毅[1,2]**,李 真[1,2],余正涛[1,2],张志坤[1,2]

(1. 昆明理工大学信息工程与自动化学院,昆明,650500;
2. 昆明理工大学智能信息处理重点实验室,昆明,650500)

**摘 要:** 研究了如何使用协作分类器(协作使用条件随机场(CRFs)和支持向量机(SVM))解决领域概念实例、属性及属性值的抽取以及它们三者之间对应关系预测的问题. 首先将概念实例、属性及属性值看作三类实体,把概念实例、属性及属性值的抽取问题转化为命名实体识别问题,利用条件随机场建模进行命名实体识别;在此基础上定义实体间对应关系,对概念实例、属性及属性值三者的对应关系做预测,把概念实例、属性与属性值三者之间存在关系的向量标记为 1,否则标记为 0,利用支持向量机建模进行关系的预测. 且以云南旅游景点概念实例、属性及属性值进行六组相关的实验. 实验表明,在开放测试中协作分类器精确度达到 84.4%、召回率达到 82.7%及 $F$ 值达到为 83.6%,相比于词语共现 $F$ 值提高了 20 个百分点.

**关键词:** 领域本体,概念实例抽取,属性抽取,属性值抽取,条件随机场,支持向量机

# Extraction and relation prediction of domain ontology concept instance, attribute and attribute value

*Guo Jian-Yi*[1,2], *Li Zhen*[1,2], *Yu Zheng-Tao*[1,2], *Zhang Zhi-Kun*[1,2]

(1. The School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China; 2. Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming, 650500, China)

**Abstract:** This paper studies how to use the Collaboration Classifier (Conditional Random Fields (CRFs) and Support Vector Machine (SVM)) to solve the extraction and relation prediction problem of ontology concept instance, attribute and attribute value. Firstly, taken concept instance, attribute and attribute value as three entities, the problem of extraction these three entities was converted to a named entity recognition problem, CRFs classifier model was adopted to recognize entities; Furthermore, made a definition for the relations between the concept instance, attribute and attribute value and made relations prediction among concept instance, attribute and

* 基金项目:国家自然科学基金(60863011),云南省自然科学基金(2008CC023),云南省中青年学术技术带头人后备人才项目(2007PY01-11),云南省教育厅基金(07Z11139)
收稿日期:2011-06-15
** 通讯联系人,E-mail:Gjade86@hotmail.com

attribute value after they were identified respectively, if there is a relationship among the concept instance, attribute and attribute value, marked 1, otherwise marked 0, then use SVM classifier model to make predictions on entity corresponding relation. Taking six trials on concept instance, attribute and attribute value on Yunnan tourist attractions for instance, the experiment is done to make that the accuracy rate of Collaborative Classifier achieves 84.4% and recall rate is up to 82.7% and the $F$ score is 83.6%, compared to Words Co-occurrence model, its $F$-score increased by 20%.

**Key words**: domain ontology, concept instance extraction, attribute extraction, attribute values extraction, conditional random fields, support vector machine.

本体是由概念和概念之间关系组织起来的结构体. 对于特定领域,领域概念(往往就是领域术语)则比较具体,包括:领域概念、实例、属性及其属性值. 概念之间的关系(instance-of, attributes-of 等)是通过概念实例和属性来表现的,同一概念下的实例具有相同的基本属性,不同的属性可以区分概念实例属于不同的概念. 由领域概念层、属性层、实例层及属性值层共同组成了领域知识本体的层级分类体系[1]. 由此可见,在本体概念关系学习过程之中,概念实例和属性是必不可少的两个部分,而属性需要用属性值来描述,因此,属性值的抽取在本体知识库应用中也具有同样重要的意义.

目前国外有关概念实例属性抽取研究较为广泛,文献[2]中 D Sánchez 提出并实现了一个自动的、无监督的且领域无关的概念实例属性抽取方法;文献[3]利用句法模式从 Web 上抽取候选概念属性,并将判别属性看作分类问题,利用两个有指导的分类器来进行分类;文献[4]使用无指导的方法从半结构化的 HTML 文档中抽取属性和属性值对. 文献[5]利用弱指导的方法从结构化的 Web 文档中抽取概念属性. 目前国内的相关研究则相对较少,文献[6]提出了一种基于 Web 弱指导的本体概念实例和属性的抽取方法,利用小规模的种子实例和属性集,从 Web 上自动获取实例和属性共现的上下文模式,并利用种子实例和属性的关联性来评价这些模式,结果达到了实用水平;文献[7]利用支持向量机抽取人的相关属性,并取得了较好的抽取结果. 但上述研究并未进行领域本体概念实例、属性及属性值的抽取以及三者之间的关系预测工作的研究. 由于概念实例就是实体,而概念属性一般是名词,属性值也往往用数字或名词来描述,因此,可将领域本体概念实例、属性及属性值这三者看做实体.

本文提出了一种基于条件随机场和支持向量机协作使用的方法,对本体概念实例、属性及属性值进行抽取,利用分类及命名实体(named entity, NE)识别的思想,对三者做实体识别,实现对领域概念实例、属性及属性值的抽取,再在此基础上做实体对应关系的抽取. 抽取之前需要对实体对应关系做定义:(1)概念实例与属性之间的对应关系(如:梅里雪山与平均海拔);(2)概念属性与属性值之间的关系(如平均海拔与 6000 m). 把概念实例、属性及属性值之间存在关系的向量标记为 1,否则标记为 0,然后利用 SVM 建模进行两两之间关系的预测. 以云南旅游景点概念、实例及属性进行试验验证,结果表明该方法具有良好的可行性.

## 1 方法原理

**1.1 协作分类思想** 本文主要研究自由文本中本体概念实例、属性及属性值的抽取. 首先将本体概念实例、属性及属性值看作实体,利用条件随机场对训练语料建模进行实体识别,但经识别后得到的三类实体不存在任何关系;因此,接下来的工作就是对实体关系预测,根据文本结构抽取特征,利用支持向量机进行建模以实

现概念实例、属性及属性值的对应关系的预测.

本文仅对出现在同一句话中的本体概念实例、属性及属性值进行抽取,对于出现在不同句子中的本体概念实例、属性及属性值,采取上下文就近匹配的原则,就是说在一句话中仅出现了本体概念实例、属性及属性值三类实体的2种或是1种,就采取回溯或是前进法在上一句(上二句)或是下一句(下二句)查找可能与之匹配的实体,回溯或前进的阈值根据旅游领域经验值设置为2.本体概念实例、属性及属性值抽取系统架构图及示例结果列表分别如图1和表1所示.

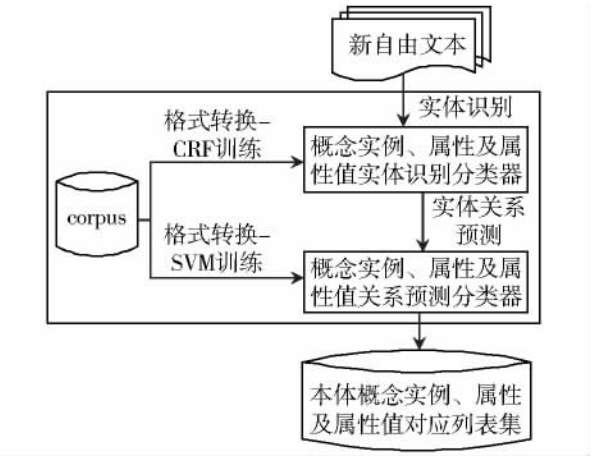


图1　概念实例、属性及属性值抽取系统架构

Fig. 1　The architecture diagram of concept instance, attribute and attribute values extraction system

表1　概念实例、属性及属性值抽取结果

Table 1　The results of concept instance, attribute and attribute value extraction

| 景点概念实例 | 属 性 | 属性值对应列表典型样例 |
|---|---|---|
| 属都湖景区 | 总面积 | 约 300 $km^2$ |
| 梅里雪山 | 平均海拔 | 6000 m 以上 |
| 金马碧鸡坊 | 位置 | 金碧路与三市街交汇处 |
| 洱海 | 蓄水量 | 28 亿立方米 |
| 滇池 | 蓄水量 | 13.7 亿立方米 |
| 云南陆军讲武堂 | 始建时间 | 明朝宣德年间 |
| 金殿公园 | 门票价格 | 20 元/人 |
| … | … | … |

## 1.2　基于条件随机场的概念实例、属性及属性值的实体识别

**1.2.1　条件随机场**　条件随机场(Conditional Random Fields, CRFs)常被用于中文分词和词性标注等词法分析工作,它没有隐马尔可夫模型那样严格的独立性假设,具有表达长距离依赖性和交叠性特征的能力,能够较好地解决标注(分类)偏置等问题的优点,而且所有特征可以进行全局归一化,能够求得全局的最优解.而实体识别问题即可定义成序列的标注,即判断观察词是否属于预先定义的特征集合,恰恰应合了条件随机场用于序列标注的优势,所以条件随机场模型非常适用于命名实体识别[8].

对于NE识别,对给定词序列 $x=(x_1,x_2,\cdots,x_n)$ 和标注序列 $y=(y_1,y_2,\cdots,y_n)$ 定义一个条件随机场模型如式(1)所示:

$$p(y\mid x)=\frac{1}{Z(x)}\exp\left(\sum_i\sum_k\lambda_k f_k(y_{i-1},x)\right) \tag{1}$$

其中 $z(x)$ 是归一化因子,$n$ 表示给定词序列的长度,$f_k(y_{i-1},y_i,x)$ 是特征函数可以更具体的讲解一下,既可以表示无向图边的转移特征 $e(y_{i-1},y_i,x)$,也可以表示节点的状态特征 $v(y_i,x,i)$,$\lambda_k$ 是第 $k$ 个特征函数的权重系数.

**1.2.2　语料的标注转换**　概念实例、属性和属性值的识别任务对定义了六种标记集合,即 $L=$(JDB, JDM, JDE, SXB, SXM, SXE, PVB, PVM, PVE, O),其中各个标记的意思分别是景点首部、景点内部、景点尾部、属性首部、属性内部、属性尾部、属性值首部、属性值内部、属性值尾部、其它.例如:"轩辕台始建于战国至西汉年间",按照标注体系将其分为观察、标注对序列:"轩辕/JDB 台/JDE 始建/SXB 于/O 战国/PVB 至/PVM 西汉/PVM 年间/PVE".

**1.2.3　特征选择与特征模版的制定**　利用条件随机场的机器学习方法进行实体识别,最重要的过程就是实体特征的选择和特征模版的制定.CRFs模型最大的优点就是不仅能够综合使用字、词、词性等上下文信息,还能综合利用各种外部特征,这个特点在内部基本特征以外

的很多方面都有尝试的空间. 通过定义模板来筛选特征,在特征模版的选取中,融合了一元特征、二元特征以及多元特征,并将词表纳入外部特征. 实验时,对于概念实例其主要选取的特征包括:词本身、词性、后缀特征词、左右指界词、常见概念实例、组合特征、上下文特征信息(活动窗口);对概念属性,主要选取的特征包括:词本身、词性、组合特征、上下文特征信息;对概念属性值,主要选取的特征包括:方位词、数量单位词、左右指界词、组合特征、上下文特征信息(活动窗口)等. 通过对训练语料和多次试验的结果分析,选取词本身、词性以及词本身和词性的多元混合特征时,试验达到最佳效果,CRFs 特征模版示例如表 2 所示.

表 2 特征模版项示例

Table 2 Characteristics of the template in sample

| 特征类型 | 模版项 | 标示内容 |
|---|---|---|
| 一元特征 | %x[1,0] | 当前行的下一行,第 0 列(词本身) |
| | %x[1,1] | 当前行的下一行,第 1 列(词性) |
| 一元复合特征 | %x[2,1]/%x[2,1] | 当前行的下二行的第 1 列和第 1 列(词性,词性) |
| | %x[2,0]/%x[2,1] | 当前行的下二行的第 0 列和第 1 列(词本身,词性) |
| 二元特征 | %x[-2,0]/%x[-1,0] | 当前行的上二行和上一行的第 0 列(词本身) |
| | %x[-2,1]/%x[-1,1] | 当前行的上二行和上一行的第 1 列(词性) |
| 多元特征 | %x[-2,1]/%x[-1,1]/%x[0,1] | 当前行的上二行、上一行及当前行第 1 列(词性) |
| | %x[-1,1]/%x[0,1]/%x[1,1] | 当前行的上一行、当前行及下一行的第 1 列(词性) |

1.2.4 模型的训练与测试 在 CRFs 模型训练过程中,常用的训练方法有迭代梯度方法,如 GIS[9] 和 IIS 方法[10]. 这两种方法实现起来比较简单,但存在收敛慢的缺点. Wallach[11] 用变化斜率方法和二阶方法结合起来训练模型,达到了较好的效果. 所以本论文采用 Wallach 提出的方法,本系统中使用了 CRF++工具包.

1.3 基于支持向量机的概念实例、属性及属性值之间实体关系预测

1.3.1 支持向量机 支持向量机的实现是通过某种事先选择的非线性映射(核函数)将输入向量映射到一个高维特征空间,在这个空间中构造最优分类超平面.

定义一个样本点到超平面的分类间隔:$\delta_i = y_i(wx_i + b)$,现在把 $w$ 和 $b$ 进行归一化,即用 $w/\|w\|$ 和 $b/\|w\|$ 分别代替原来的 $w$ 和 $b$,那么间隔就可以写成:$\delta_i = \frac{1}{\|w\|}|g(x_i)|$,叫做几何间隔,几何间隔所表示的正是点到超平面的距离,几何间隔与 $\|w\|$ 成反比,因此最大化间隔就是最小 $\|w\|$,求解最优分类面就是寻找最小化 $\|w\|$.

1.3.2 特征选择 通过对训练语料的分析,发现在领域文本中,景点概念实例、属性以及其属性值多在同一句话出现,例如:"滇池位于昆明市西南面,海拔 1886 m,湖面南北长 39 km,东西宽 13.5 km,平均宽度约 8 km". 所以在构造实体对时,采取以下原则:将出现在同一句话中的实体,进行"概念实例-属性"及"属性-属性值"配对. 同时对训练语料句子结构分析,可以发现景点概念实例、属性及属性值左、中、右的 $k$ 个词及词性对确定它们间的对应关系有很好的区分作用,此外概念实例和属性以及属性和属性值的出现顺序和距离也是决定它们之间关系的重要因素. 此外,由于标注的语料数量有限,相对整个互联网很难覆盖所有语言情况,而且汉语的表达方式灵活多样,描述同一概念可能用多个词,例如,描述地理位置相关的特征词汇就可能有"位于"、"坐落在"、"地处"、"距离"、"距"等,字面描述的差异造成了关键性特征的分散,使特征向量无法有效起到标识作用. 为解

决该问题,引入语义资源哈尔滨工业大学信息检索研究室的《同义词词林扩展版》[12],把所有词都转化为其同义词在同义词词林中第 1 个词. SVM 特征模版示例项如表 3 所示.

表 3　SVM 分类器特征选择

Table 3　SVM feature selection

| 特征集编号 | 特征选取 |
|---|---|
| 1 | 实体对中两实体所属实体类型及其前后两词所属类型 |
| 2 | 实体对实体的出现顺序及其间隔距离 |
| 3 | 实体对实体左、中、右的 2 个词(转化为同义词林第一个词) |

1.3.3　模型的训练与测试　在 SVM 模型训练过程中,采用学习能力和能力较强的混合函数,通过有效特征选择和训练参数的调整得到了良好的关系预测效果[13]. 本系统中使用了 libsvm2.9 工具包.

1.4　词语共现　词语共现在信息检索、交叉销售、自然语言处理等领域是一种十分重要的关联挖掘技术[14,15]. 该技术通过分析经常在一起搭配出现的对象来分析对象之间的关联程度. 在自然语言处理中,在提供一个种子词的情况下,通过分析其他和种子词经常搭配出现的词语,就可以获得和种子词关联程度比较大的词语. 基于这种思想,在本文中也尝试用词语共现来做概念实例、属性和属性值的抽取,以和上述方法作对比[16]. 概念的实例往往和其属性及属性值同时出现,本文利用少量的种子实例、种子属性和种子属性值,抽取实例和属性、属性和属性值之间共现的上下文模式,并进一步利用上下文模式来识别概念实例、属性和属性值三者之间的关系.

## 2　试验数据及结果分析

2.1　试验数据及评价指标　根据测试集和训练集的不同关系,可以将评测分为封闭测试和开放测试,封闭测试即用已经标注的语料对领域概念实例、属性及属性值的抽取及关系的预测进行测试,开放测试即用未标记的生语料对概念实例、属性及属性值的抽取及关系的预测进行测试. 为了能够客观评价协作使用条件随机场和支持向量机在领域概念实例、属性和属性值的抽取问题上具有良好的可行性,分别做了六组实验,其中前两组实验是基于条件随机场模型进行概念实例、属性及属性值实体识别的封闭测试与开放测试;中间两组是基于支持向量机进行概念实例、属性及属性值对应关系预测的封闭测试与开放测试;后面两组是协作分类器与词语共现的比较. 目前国内有关属性抽取的研究相对较少,因而还没有符合开展此项研究要求的权威语料,于是自行构建了旅游领域语料库. 从互联网收集了 1500 篇关于云南旅游景点介绍的文本,充分利用了 Web 信息的冗余性,有效地克服了从单一文本中进行信息抽取带来的数据稀疏问题,并分别对概念实例、属性及属性值进行了标注,建立了针对云南旅游景点类概念实例、属性及属性值抽取的语料集,其中以 1000 篇作为训练语料,其余 500 篇作为测试语料. 本文采用识别的准确率、召回率和 $F$ 值作为最终的评价标准.

$$\text{准确率}(P)=\frac{\text{正确标注的实体及个数}}{\text{标注实体的总个数}}\times 100\%$$

$$\text{召回率}(R)=\frac{\text{正确标注的实体及个数}}{\text{标注实体的总个数}}\times 100\%$$

$$F\text{ 值}(F-\text{score})=\frac{2\times P\times R}{P+R}\times 100\%$$

2.2　结果及分析　通过对表 4～6 试验结果的比较分析,在领域概念实例、属性及属性值实体识别分类器训练测试过程中发现在选取实体词词本身、词性、复合特征、二元特征、多元特征及窗口浮动大小为 2 时能取得较好的抽取效果;在概念实例、属性及属性值对应关系预测分类器训练测试过程中发现,在只选取实体对各实体词本身、词性以及各实体词前后 2 个词作为特征时,效果并不理想,但加入实体对实体词间隔距离和其出现的先后顺序时,预测结果有了大大的改进.

表 4 利用概念实例、属性及属性值实体识别分类器测试结果

Table 4 The results of concept instance, attribute and attribute values recognition

| 分类器 | 开放/封闭 | 实体总数 | 识别个数 | 正确个数 | 正确率(%) | 召回率(%) | F 值(%) |
|---|---|---|---|---|---|---|---|
| 条件随机场 | 封闭测试 | 6888 | 6839 | 6750 | 98.70 | 98.00 | 98.35 |
| 分类器 | 开放测试 | 3563 | 3534 | 3130 | 88.57 | 87.85 | 88.21 |

表 5 利用概念实例、属性及属性值对应关系预测分类器试验结果

Table 5 The results of relationship prediction of concept instance, attibute and attribute value

| 分类器 | 开放/封闭 | 存在关系总数 | 识别个数 | 正确个数 | 正确率(%) | 召回率(%) | F 值(%) |
|---|---|---|---|---|---|---|---|
| 支持向量机 | 封闭测试 | 2188 | 2268 | 2131 | 93.96 | 97.40 | 95.65 |
| 分类器 | 开放测试 | 1324 | 1337 | 1208 | 90.35 | 91.24 | 90.79 |

表 6 概念实例、属性及属性值协作分类器与词语共现实验抽取结果比较

Table 6 The comparison results of concept instance, attribute and attribute value collaboration classifier with word co-occurrence

| 分类器 | 测试类型 | 实体总数 | 识别个数 | 正确个数 | 正确率(%) | 召回率(%) | F 值(%) |
|---|---|---|---|---|---|---|---|
| 协作分类器 | 开放测试 | 1437 | 1409 | 1189 | 84.39 | 82.73 | 83.55 |
| 词语共现 | 开放测试 | 1437 | 1631 | 962 | 58.98 | 67.18 | 62.81 |

另外从表 6 可以看出,协作分类器在进行领域概念实例、属性及属性值对应关系预测时,比较单分类器进行概念实例、属性及属性值对应关系预测的效果有所下降,主要原因是由于概念实例、属性及属性值实体识别分类器识别结果误差所致,一些错误噪音遗留到概念实例、属性及属性值关系预测文本中,因此在进行实体配对与特征选择上产生了错误;再者也可从概念实例、属性及属性值协作分类器与词语共现实验抽取比较,结果看出词语共现的方法不适合做领域概念实例、属性及属性值的关系抽取任务,主要是因为汉语词中有大量修饰词、名词及停用词等,诸如:风光、景观、旁边、动物、全国、人们、人民等等,这些词不但没有对信息抽取带来任何意义,还加入了很大噪音干扰. 另外从抽取结果数据中,亦可以看出在开放测试与封闭测试还存在一定的差距,这主要是由于测试语料覆盖不够全面,特征选择不够精确导致.

## 3 结 论

领域本体概念实例、属性及属性值的抽取是近年来文本信息处理领域的一个研究热点和难点. 本文介绍了如何协作使用条件随机场和支持向量机两种机器学习方法从自由文本中抽取景点类概念实例、属性及属性值,并对这三类实体进行关系预测. 通过在旅游领域对景点概念实例、属性及属性值试验,结果表明该方法具有良好的效果. 同时,领域本体概念实例、属性及属性值的抽取对领域本体构建中新术语的发现及领域本体构建中术语关系的发现都有着指导性的意义.

### References

[1] Eric T, Wang W M. A concept-relationship acquisition and inference approach for hierarchical taxonomy construction from tags. Information

Processing and Management: An International Journal, 2010, 46(1): 44～57.

[2] Sánchez D. A methodology to learn ontological attributes from the Web. Data and Knowledge Engineering, 2010, 6(69):57～597.

[3] Poesio M, Almuhareb A. Identifying concept attributes using a classifier. Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition, Ann Arbor, 2005,18～27.

[4] Yoshinaga N, Torisawa K. Open-domain attribute-value acquisition from semi-Structured texts. Proceedings of the OntoLex 2007,Busan, South-Korea, 2007, 55～66.

[5] Ravi S, Pasca M. Using structured text for large-scale attribute extraction. Proceedings of the 17th International Conference on Information and Knowledge Management. Napa Valley, California, USA, 2008, 1183～1192.

[6] Kang W, Sui Z F. Ontology concept instances and attributes simultaneously extracted based on web. Journal of Chinese Information Processing, 2010,1:54～59.(**康　为, 穗志方. 基于Web弱指导的本体概念实例及属性的同步提取. 中文信息学报**,2010,1:54～59).

[7] Ye Z, Lin H F, Su S, *et al*. Extraction of character attributes based on support vector machine. Computer Research and Development, 2007, 2:271～275.(**叶　正, 林鸿飞, 苏　绥等. 基于支持向量机的人物属性抽取. 计算机研究与发展**, 2007,2:271～275).

[8] Guo J Y, Xue Z S, Yu Z T, *et al*. Named entity recognition based on cascaded conditional random fields. Journal of Chinese Information Processing, 2009,5:47～52.(**郭剑毅, 薛征山, 余正涛等. 基于层叠条件随机场的旅游领域命名实体识别. 中文信息学报**, 2009,5:47～52).

[9] Darroch J,Lauritzen S,Speed T. Markov fields and log-linear interaction models for contingency tables. Annals of Statistics, 1980, 8(3): 522～539.

[10] Della P S, Della P V, Lafferty J. Inducting features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(4):380～393.

[11] Wallach H. Efficient Training of conditional random fields. http: www. cogsci. ed. ac. uk, 2002.

[12] Information Retrieval Laboratory, Harbin Institute of Technology. Synonymous with the word forest (Extended Edition). http://www.ir-lab.org/, 2008-05-19.(**哈尔滨工业大学信息检索研究室. 同义词词林(扩展版)**. http://www.ir-lab.org/, 2008-05-19).

[13] Liao S Z, Ding L Z, Jia L. Support vector regression parameter adjustment. Journal of Nanjing University (Natural Sciences), 2009, 45(5):585～592.(**廖士中,丁立中,贾　磊. 支持向量回归多参数的同时调节. 南京大学学报(自然科学)**,2009, 45(5):585～592).

[14] Geng Q, Geng C. Use of the word co-occurrence for Ontology concept gain. Modern Library and Information Technology, 2006,1(2): 43～45.(**耿　骞,耿　崇. 利用词语共现进行Ontology的概念获取. 现代图书情报技术**, 2006,1(2):43～45).

[15] Geng H T, Cai Q S, Yu K, *et al*. Document keywords automatically extracted based on word co-occurrence map. Journal of Nanjing University (Natural Sciences), 2006, 42(2):156～162.(**耿焕同,蔡庆生,于　琨等. 一种基于词共现图的文档主题词自动抽取方法. 南京大学学报(自然科学)**,2006, 42(2):156～162).

[16] Yao X M Guo J Y Yu Z T, *et al*. A new algorithm based on word co-occurrence and its application in domain concept extraction. 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems, Shanghai, China, 2009, 4(3):521～525.