文章编号:1001 - 9081(2011)08 - 02138 - 05
doi:10.3724/SP.J.1087.2011.02138

基于领域本体的专业文档语义标注方法

魏墨济,于　涛
(中国海洋大学 信息科学与工程学院,山东 青岛 266100)
(weimoji@126.com)

摘　要:提出了一种标注方法实现语义网中无结构专业文档的自动标注。通过分析给出专业文档的两方面特征,并提出了三个假设。为提高标注效率,基于结构对本体进行分割,将本体划分成具有较高语义独立性的片断;然后利用从专业文档中抽取的关键词定位本体片断;最后,使用选中的片断,利用语法结构和三元组的对应性对文档进行标注。实验结果表明,所提出方法在标注效率、标注数量和准确性三个方面都有所提高。

关键词:专业文档;语义标注;本体分割;语义环境;语法环境
中图分类号: TP182; TP391.3　　**文献标志码**:A

Professional literature annotation method based on domain ontology

WEI Mo-ji, YU Tao
(*College of Information Science and Engineering, Ocean University of China, Qingdao Shandong 266100, China*)

Abstract: An automatic annotation method for professional literature was proposed. Through comparing with other storage formats and literary styles, two features of professional literature were summarized, and then three assumptions were proposed. To improve annotation efficiency, based on topology structure, the domain ontology was partitioned into segments which were self-consistent, then the most related segments were located with the keywords extracted from document, finally the document with located segments was annotated and the annotation scope was expanded according to the correspondence between grammatical structure and semantic structure. The experimental results show that the proposed method can improve annotation efficiency, annotation quantity and annotation accuracy.

Key words: professional literature; semantic annotation; ontology partition; semantic context; syntax context

0　引言

各种类型的专业文献资料是人们获取知识的重要手段和途径。这些文档一般由领域专家撰写,具有较高的可信度,是人们进行学习和研究的基础资源。为提高专业文献资料的共享程度和传播速度,各领域纷纷将专业文档电子化并发布到网络上,以方便人们的检索。然而当前基于语法关键词匹配的搜索方法很难解决一词多义和多词一义所带来的歧义问题,因此人们为了获取某知识通常需要多次搜索该知识相关的关键词,并手工地过滤掉一词多义所带来的冗余信息。这种检索方法将花费人们大量的时间和精力去检索资源,其效率非常低下。随着网络文档的极大丰富,有效资源的检索已成为知识获取的瓶颈。

通过对当前万维网进行语义扩展,人们提出了语义网[1]。语义网为知识查找提供了一种基于语义的检索方法。它利用形式化的本体对网络上的各种资源进行标注,使得标注后的资源更加适合机器处理[2]。在语义检索中计算机根据语义对资源和关键词的相关度进行预判断,以提高检索的效率,节省时间。该检索方法的前提是网络中的资源已得到较好的标注。然而相对于语义网的其他技术,语义标注技术仍显得较不成熟。近年来自动语义标注得到越来越多研究团体的重视,但在各种网络资源的标注中,仍很少有研究涉及专业文档的标注。

当前已有对网络中文本内容进行标注的研究,但由于网络语言和格式的多样性,仍没有一种能够满足所有情形的通用标注方法,不同文献中的标注方法各有不同特色,所针对的标注对象也不尽相同。文献[3]利用已有的语义指针,通过本体中的概念关系和词汇间的同义关系对未知的词汇进行注释;文献[4]为了使本体能够与应用程序有效地集成在一起,将本体的标注作为服务提供给用户,这与通常所采用的文件的形式提供给用户不同;文献[5]使用概念识别工具和命名实体识别工具来标注生物学内容;文献[6]使用多个不同本体对文本内容进行注释,当这多个不同本体出现不一致时,使用 hierarchical free tagging 方法来统一概念;文献[7]提出了一种树结构的条件概率模型,来研究网页上需要标注的目标实例,使得标注过的实例之间存在层次关系或兄弟关系;文献[8]对使用了大量日常用语、缩写、俚语等词汇的非专业文档的标注方法进行了研究;文献[9]利用 Relation Population 和 Predicate Suggestion 对维基百科网页上的内容进行了标注;文献[10]通过对 state-of-the-art positive-only learning 算法进行扩展,提出了 B-POL 算法自动从维基百科中抽取关系;文献[11]使用用户桌面中的信息对网页进行标注,使得这些注释过的网页具有个性化的特征,更加符合用户的使用习惯;文献[12]研究用户的标记和分类学之间的映射,对用户的标记进行注释。

此外在图书馆分类学中也有对文献资料进行标注的研究[13-16],通过本体语义标注可以实现文献的分类管理,对文献领域的划分具有指导意义。

收稿日期:2011 - 01 - 24;**修回日期**:2011 - 04 - 20。

作者简介:魏墨济(1981 -),男,山东济南人,博士研究生,主要研究方向:语义网、本体、Web Service;　于涛(1984 -),男,山东青岛人,硕士研究生,主要研究方向:语义网、本体。

1　专业文档分析

1.1　专业文档特征

本文研究的标注对象是领域专家所书写的专业文档。在给出标注方法之前,本文先对其特征进行分析。

首先专业文档一般采用 txt、doc、pdf 等文本文件的格式存储,此格式文档与网页或数据库在存储结构方面有明显的不同。网页是一种半结构化的文档,数据库是一种结构化的文档,在这种半结构化或结构化的文档中可以采用 XML 或 E-R 图来说明概念以及概念间的关系。而专业文档是一个无结构的文档,其中的概念以及概念间的关系都隐含在文档之中而没有显性的说明。当前语义网相关的标注方法大都是基于文档的结构实现的,这些方法很难应用于无结构的专业文档。因此本文从专业文档的语法结构入手,找出隐含在文档中的概念并将其关系显性化。

其次专业文档用于领域知识的说明,因此其具有不同于小说、诗歌、散文等其他体裁文档的特点。专业文档更倾向于对领域中的概念以及概念间的关系进行说明,因此专业文档的用词更为规范,内聚度也更高。在图书馆分类学中一般是利用轻量级本体或分类学元数据对文档进行标注,该标注方法能够帮助对专业文档所属领域进行识别。但相同领域的专业文档所涉及的概念可能大不相同,仅使用标题、关键词、摘要等少量信息进行标注,很难清楚地表达文档所要描述的概念。因此基于轻量级本体和分类学元数据的标注很难实现专业文档中概念的精准定位。当前各个领域中大都已有较为通用的领域本体。如地理信息领域的 GEO-Ontology;基因领域的 Gene Ontology 等。这些重量级本体涵盖了所描述领域的重要概念,且其语义信息更为丰富,这些本体已得到领域专家的广泛认可,具有较高学术价值和应用价值。本文可以使用这些成熟的重量级本体来标注专业文档的概念。

通过对专业文档存储格式和体裁的分析可以看出专业文档具有两方面的特点:1)无结构,概念及概念间的关系没有显性的说明;2)文档中的概念密集。专业文档这两方面的特点使得当前已有的标注方法很难应用于专业文档标注。为了标注专业文档本文对其内容和用词进行了深入研究,并提出了 3 个假设。

1.2　专业文档的 3 个假设

假设 1　文档是顺序结构的,文档的每一部分都是对领域子域的解释,也即文档对应领域本体的局部区域。

每个领域都包含着丰富的内容,涵盖了大量的概念,如 NCI Ontology 中拥有 17000 多个概念。而一篇专业文档不可能说明整个领域,只能是对其中部分概念和属性进行说明。

假设 2　文档的各个部分之间是有关系的,部分之间的连接是自然的而非突兀的。

专业文档的各部分之间具有较强的相关性,也就是说专业文档所涉及的概念在语义上是有联系的。

假设 3　同一文档中所使用的词汇是一致的。

也就是说同一文档中的相同词汇其语义是不变的,而同一文档中为表达同一含义所使用的词汇也是不变的。词汇与语义是一一对应的。

为清楚地说明词汇和其所处文档部分之间的关系,本文给出语法环境的定义。

定义 1　语法环境。令 D 是一篇文档,w 是文档中的词汇,P 是文档中的段落,Ch 是文档中的章节。当 $w \in P, P \subseteq Ch$ 并且 $Ch \subseteq D$,称 P、Ch、D 为 w 的语法环境。其中 D 为 w 的最大语法环境,P 为 w 的最小语法环境,Ch 为 w 的中间语法环境。

2　本体分割

由假设 1 可知任意一篇专业文档只能是对领域本体中的部分概念进行描述,而领域本体通常比较庞大,因此若使用整个本体对文档进行标注,会有大量的计算资源浪费在本体的解析和概念的查找上,这势必降低标注的效率。如文档中的词汇 models 与 ACM(Association for Computing Machinery)本体中的 Data Models、Process Models、Language Models 等多个概念相关,这些概念之间具有较小的语法距离和较大的语义距离。当上下文语义模糊时,对 models 进行标注时很容易产生错误。通过将语义距离较远的概念划分到不同本体片断的方式,来控制语法相关度较高词汇之间的干扰。由假设 2 可知专业文档所描述的概念在语义上是有联系的。在本体的形式化定义中,属性拥有 Domain 和 Range 两个要素,这两个要素分别说明了属性的定义域和值域,将两个相关的概念联系在一起,并对两个概念的关系进行说明和约束。通过属性的连接,本体中语义关系密切的概念其空间距离将非常接近,随着语义关系的疏远其空间距离也将不断增大。为提高标注的效率和准确率,本文对本体进行分割使用本体片断对文档进行标注。

当前对本体的分割主要有两种方式:一是根据本体的逻辑结构分割;二是根据本体的拓扑结构分割。第一种分割方式需要根据公理或属性计算每对概念之间的相关程度,随着概念个数的增加其计算量呈指数增长,因此该分割方式的算法复杂度极高,并且在对较大本体进行分割时常常失败。通常领域本体往往具有众多的概念,因此基于逻辑结构的分割方式很难在有效的时间内对重量级本体进行有效地分割。而本文所要关注的是关系密切的概念在空间距离上的关系,因此采用基于拓扑结构的方式分割领域本体。

Stuckenschmidt 等人[17-20]研究了如何基于结构对大本体进行分割,他们将分割过程分为五个步骤:创建关系图、计算关系强度、划分模块、分配孤立概念和合并小模块。他们将一个概念划分到一个片断中当且仅当它与片断中概念的关系强于与片断外概念的关系,因此该方法能将拥有较多连接的概念划分到一个片断中,所划分的片断具有较高的语义独立性。此外他们还给出了分割工具 Pato。在本文的标注研究中,也采用 Pato 工具对本体进行分割。

为清楚地说明词汇和本体片断之间的关系,本文给出语义环境的定义。

定义 2　语义环境。O 是领域本体,Os 是本体片断,C 是 Os 中的概念,w 是词汇。如果 $C \in Os, Os \subseteq O$ 并且 $w^I \in C^I$,称 Os 为 w 的语义环境。

3　专业文档标注

3.1　本体片断选择

在对领域本体有效分割后,便可以使用本体片断对专业文档进行标注了。在标注之前首先需要为文档选择语义相关度最高的本体片断,其选择过程如下:

1)在最大语法环境中利用本体学习技术查找关键词汇。

2)通过关键词与本体片断中概念的对比,为语法环境匹配相关度最高的语义环境。

3)如果关键词所对应的概念集中在同一个语义环境中,则匹配成功,使用匹配的语义环境标注语法环境。

4)反之,如果关键词所对应的概念均匀地分散在多个语义环境中,则匹配失败。

5)当匹配失败时,缩小语法环境,在缩小的语法环境中查找关键词汇,重新与语义环境匹配。重复3)和4),直到匹配成功或缩小到最小语法环境。

本体片断选择算法如算法1所示。

算法1 Segments selection。

```
Boolean isMinContext = False;
Ontology Segment OS, MOS[];
SytaxEnviromentStack stack;
SytaxContext curSyCon;
INPUT Ontology Segments;
INPUT document;
push stack(document);                    // push document to the stack
DO WHILE NOT empty stack
    curSyCon = pop stack();
    Extract keywords from curSyCon;
    IF curSyCon equals minimum syntax context
        isMinContext = TRUE;
    ELSE
        isMinContext = False;
    ENDIF
    //the if condition is determined by algorithm 2
    IF keywords-related concepts evenly scattered to several semantic
        contexts
        IF NOT isMinContext
            shrink syntax context;
            push stack(shrunk syntax context);
        ELSE
            MOS[] = multi semantic contexts;
            //MOS[] is MSC[] in algorithm 2
            annotatetransitionalparagraph(curSyEnv, MOS);
        ENDIF
    ELSE
        OS = single semantic context;
        // OS is SC in algorithm 2
        annotate(curSyEnv, OS);
    ENDIF
ENDDO
```

在关键词抽取过程中本文利用本体学习技术计算词汇与文档所属领域的相关度。其计算公式如式(1)所示:

$$DR_{t,k} = \frac{P(t \mid D_k)}{\max\limits_{1 \leqslant i \leqslant n} P(t \mid D_i)} \tag{1}$$

其中:D_k 是领域集合 $DS = \{D_1, D_2, \cdots, D_n\}$ 中文档所属的领域;t 是从文档中抽取出的词汇,词汇 t 与领域 D_k 的相关度用 $DR_{t,k}$ 表示,条件概率 $P(t \mid D_k)$ 可用式(2)估算:

$$E(P(t \mid D_k)) = f_{t,k} / \sum_{t \in D_k} f_{t,k} \tag{2}$$

其中 $f_{t,k}$ 表示词汇 t 在领域 D_k 中出现的频率。

本文取相关度较高的若干词汇作为关键词来定位语义环境。若使用语义距离计算的方法,需要计算每一个概念与关键词之间的相关度,虽然可以准确地找到与关键词匹配度最高的概念来定位语义环境。但如前所述领域本体含有大量的概念,其定位效率将会比较低。因此本文采用语法匹配加统计的方法来定位语义环境。定位步骤如下:

1)使用字符串匹配的方法,检查每一个本体片断是否含有关键词相似的概念,如果有则称本体片断命中关键词,并在本体片断中记录关键词。

2)统计每个本体片断所命中关键词的个数,并计算每个本体片断的命中率。

命中率 = 命中关键词的个数/关键词总数

3)比较命中率。

①若某本体片断的命中率远大于其他本体片断的命中率,则语义环境定位成功。

②若多个本体片断的命中率相近,则语义环境定位失败。失败原因又可分为两种情形。

情形1 如果命中率相近的本体片断所记录的关键词重复率很低,说明所选语法环境不合适。缩小语法环境,利用新选出的关键词重新定位语义环境。

情形2 如果命中率相近的本体片断所记录的关键词重复率较高,说明所选关键词数量不合适,则增加或减少关键词数量重新计算命中率。

在标注实验中发现,当选取的关键词较少时,会有多个本体片断的命中率相近,随着关键词数量的增大命中率差距逐渐增大,并在某个数量点处达到最大。之后随着关键词的增多命中率差距逐渐减小。当关键词过少或过多时均会出现情形2所描述的问题,因此可以采用增加或减少关键词的方法处理情形2。

算法2给出了语义环境的定位算法。

算法2 Location。

```
Integer matched[];
Sematic Contexts SC, MSC[];
String matchedwords[][];
Double matchedrate[];
Boolean finished = FALSE, sign = FALSE, increase = TRUE;
INPUT maxKeywordnum;
INPUT sematicContexts[];
DO WHILE NOT finished
    INPUT keyword[maxKeywordnum];
    FOR i = 1 TO count of semanticContexts[]
        FOR j = 1 TO maxKeywordnum
            IF StringMatch(keyword[j], sematicContexts[i])
                matchedwords[i][k++] = keyword[j];
                matched[i] ++;
            ENDIF
        ENDFOR
    ENDFOR
    FOR i = 1 TO count of semanticContexts[]
        matchedrate[i] = matched[i]/maxKeywordnum;
    ENDFOR
    IF max matched rate >> other matched rate
        SC = max matched rate sematic context;
        finished = TRUE;
        RETURN FALSE;
    ELSE
        IF NOT highoverlap(matchedwords[] which have similar hit
            rate)
            MSC = ontology segments which hit rates are close;
            finished = TRUE;
            RETURN TRUE;
        ELSE
            IF gaps among hit rates increase
                maxKeywordnum = maxKeywordnum + increasement;
                increase = TURE;
            ELSE
                maxKeywordnum = maxKeywordnum - increasement;
                increase = FALSE;
            ENDIF
```

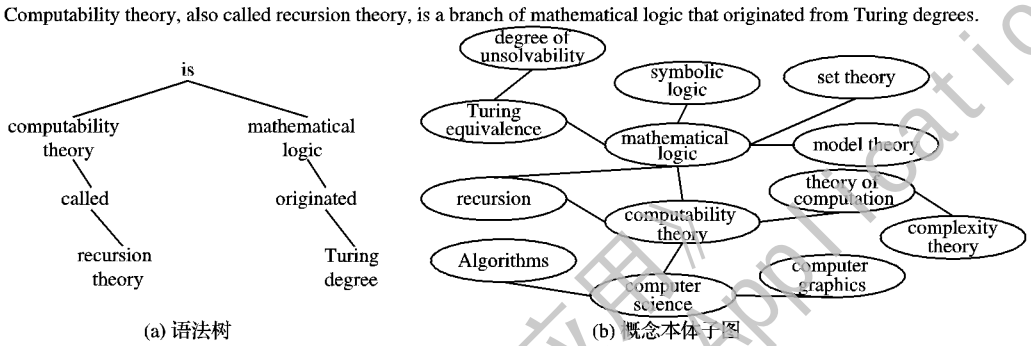
```
            clear matchedwords[ ][ ];
            IF sign XOR increase
              sign = sign XOR increase;
              increasement = increasement/2;
            ENDIF
          ENDIF
        ENDIF
      ENDDO
```

3.2　单语义环境标注

在标注实验中利用算法 1 可以为语法环境找到相应的语义环境,并且大多数语法环境仅有唯一的语义环境与其相对应,只有少部分的语法环境对应于多个语义环境。本文首先研究单一语义环境的标注问题。

如前所述,专业文档是一种无结构文档,没有标签可以显性地说明概念之间的关系。因此需要将隐性的关系显性化。文档中句子的语法结构说明了句子各语法成分之间的关系,因此可以利用句子的语法结构来显性化关系。

语义标注是为文本中的命名实体分配语义描述[21]。因此本文主要关心句子实词之间的关系。虽然对文档的每句话进行语法分析将花费较多的时间,但专业文档具有概念密集的特征,为清晰地说明概念间的关系,仍有必要对句子的语法进行解析。由假设 2 可知句子的语法成分在语义上是有关系的。本体使用属性来描述概念之间的语义关系。由此可以看出句子的语法结构和本体的三元组具有较好的对应性。因此可以利用语法结构与三元组的对应性对文档进行标注。本文以从计算理论相关的文档中抽取出的样句为例说明标注流程,如图 1 所示。



(a) 语法树　　(b) 概念本体子图

图 1　标注流程

首先使用语义环境标注关键词。通过对文档词汇的统计可以知,computability theory 是该文档的关键词,因此先用语义环境中的概念 computability theory 标注该关键词。

其次分析含有关键词句子的语法结构。采用语法树的形式描述句子各实词成分之间的关系,如图 1(a)所示。

再次在语法树中找距离已标注词汇最近的语法成分,然后使用已标注词汇所对应的概念的邻近概念标注。重复这个过程直至标完句子的所有成分。如 computability theory 的最近成分是 recursion theory 和 mathematical logic,对比它们与 computability theory 概念邻近概念的语义距离,分别使用相关度最高的 recursion theory 和 mathematical logic 进行标注。此后使用 Turing degrees 对 Turing degrees 标注。

最后根据假设 3,将概念标注扩展到其他语法环境的句子中,并解析这些句子的语法结构重复上一步标注过程。根据假设 3 可以认为文档中 mathematical theory、recursion theory 等词汇代表的语义不变,因此可直接使用 mathematical theory、recursion theory 等概念对其他句子中的相同词汇进行标注。

根据语法结构与语义结构的对应性进行标注,可以将标注的范围限定在一个较小的局域,以进一步减少参与比较的概念数量,提高标注效率。利用图结构的传递性,并时刻保持语义环境同语法环境的严格对应,逐步扩展标注范围以提高标注准确率。最后根据假设 3 快速标注相同词汇。

3.3　多语义环境标注

在标注实验中,会有少部分的最小语法环境跨越多个语义环境。当一个语法环境仅对应于一个语义环境时,则该语法环境所表达的核心观点较为单一。当一个语法环境缩小到最小语法环境后仍跨越了多个语义环境,可以认为该段落描述的知识较为分散,其意图在于说明各知识之间的关系,是一个过渡段落起到衔接和承上启下的作用,其涉及的概念也大多处于多个语义环境的交界处。因此对于这种语法环境本文主要使用语义环境间的边界概念以及连接边界概念的关系进行标注。在对此类语法环境标注之前,首先寻找连接多个语义环境的概念和属性,然后使用这些概念标注关键词汇,最后以此为基础根据语法结构和语义结构的对应性扩展标注。

4　实验及结果分析

实验中采用 ACM 本体对计算机领域专业文档进行标注测试。利用本体分割工具 Pato,将 ACM 本体划分成 21 个片断。并在网上找了 8 篇计算机领域的专业文档,其中计算机图形相关文档 2 篇、计算理论相关文档 3 篇、人工智能相关文档 3 篇。在这 8 篇文档中 4 篇来自于 Wiki 百科,2 篇来自于期刊,2 篇来自于电子图书。本文主要从标注效率和标注准确率两个方面验证所提出的方法。

4.1　标注效率

采用本文的方法标注文档所花费的时间与不采用所花费的时间,如图 2 所示,可以看出在对同一篇文档进行标注时本文的方法提高了标注效率。

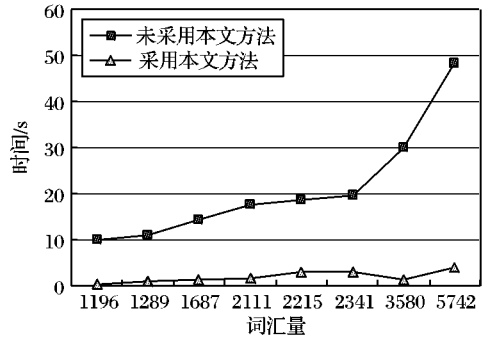


图 2　标注时间比较

4.2　标注准确性

图 3、4 分别为未采用本文方法与采用本文的方法进行标

注时的标注数量、准确率比较。通过对比可以看出在对同一篇文档进行标注时,本文的方法在标注数量和标注准确率两方面都有所提高。

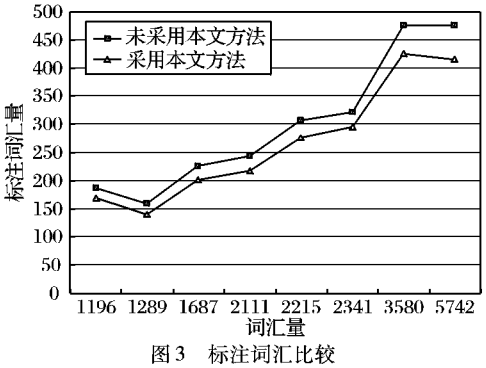


图3 标注词汇比较

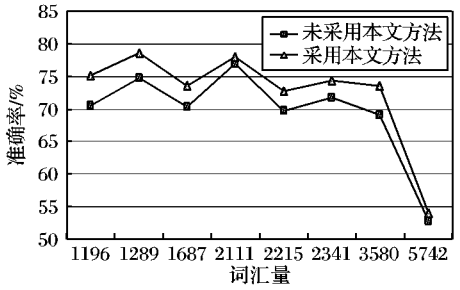


图4 标注准确率比较

5 结语

网络资源语义的缺乏已成为阻碍语义网发展的最大障碍,本文针对专业文档提出了一种自动标注方法。专业文档具有无结构和概念密集两个特征,通过对文档的分析提出了三个假设。在标注方法中,为提高标注准确率依据拓扑结构将本体分割成具有较高语义独立性的片断。利用本体学习技术从文档语法环境中抽取关键词,利用关键词为语法环境定位语义环境。使用语义环境中的概念标注关键词,并以此为基础利用语法结构与语义结构的对应性扩展标注范围。通过ACM本体对计算机领域专业文档的标注实验结果显示,所提出的方法能够增加标注的数量并提高标注的效率和准确性。

同语义网中其他资源的标注一样,也面临着海量文档的标注问题。本文仅考虑了单个文档的标注,下一步的工作主要集中在如何有效地将标注任务分布到多个网络节点上并行处理,设计一种合适的任务划分方案,使多个节点能够有效地均衡负载。

参考文献:

[1] BERNERS-LEE T, HENDLER J, LASSILA O. The semantic Web [EB/OL]. [2011-01-20]. http://old.hki.uni-koeln.de/temp/SemWebSeminal.pdf.

[2] DILL S, EIRON N, GIBSON D, *et al.* SemTag and seeker: Bootstrapping the semantic Web via automated semantic annotation [C]// Proceedings of the 12th International World Wide Web Conference. New York: ACM Press, 2003: 178-186.

[3] PAZIENZA M, STELLATO A. An environment for semi-automatic annotation of ontological knowledge with linguistic content [C]// Proceedings of the 3rd European Semantic Web Conference. Budva: Montenegro: [s.n.], 2006: 442-456.

[4] VILJANEN K, TUOMINEN J, HYVNEN E, *et al.* Extending content management systems with ontological annotation capabilities [EB/OL]. [2011-01-11]. http://www.seco.tkk.fi/publications/2007/viljanen-tuominen-hyvonen-et-al-extending-content-management-systems-with-ontological-annotation-capabilities.pdf.

[5] JONQUET C, SHAH N, YOUN C, *et al.* NCBO: Annotator semantic annotation of biomedical data [EB/OL]. [2011-01-20]. http://kcap09.stanford.edu/share/posterDemos/171/paper171.pdf.

[6] GONZÁLEZ M, BIANCHI S, VERCELLI G. Semantic framework for complex knowledge domains [EB/OL]. [2011-01-25]. http://ceur-ws.org/Vol-401/iswc2008pd_submission_17.pdf.

[7] TANG JIE, HONG MINGCAI, LI JUANZI, *et al.* Tree-structured conditional random fields for semantic annotation [C]// ISWC 2006: Proceedings of the 5th International Semantic Web Conference, LNCS 4273. Berlin: Springer-Verlag, 2006: 640-653.

[8] GRUHL D, NAGARAJAN M, PIEPER J, *et al.* Context and domain knowledge enhanced entity spotting in informal text [C]// Proceedings of the 8th International Semantic Web Conference. Washington DC, USA: [s.n.], 2009: 260-276.

[9] WANG HAOFEN, FU LINYUN, YU YONG. Bricking semantic Wikipedia by relation population and predicate suggestion [EB/OL]. [2011-01-11]. http://iws.seu.edu.cn/csws2009/paper/p97.pdf

[10] WANG GANG, YU YONG, ZHU HAIPING. PORE: Positive-only relation extraction from Wikipedia text [C]// Proceedings of the 6th International and 2nd Asian Semantic Web Conference. Berlin: Springer-Verlag, 2007: 580-594.

[11] CHIRITA P A, COSTACHE S, NEJDL W, *et al.* P-TAG: Large scale automatic generation of personalized annotation tags for the Web [C]// Proceedings of the 16th International Conference on World Wide Web. New York, USA: ACM Press, 2007: 845-854.

[12] WARTENA C, BRUSSEE R. Instanced-based mapping between Thesauri and Folksonomies [C]// Proceedings of the 7th International Semantic Web Conference. Karlsruhe, Germany: [s.n.], 2008: 356-370.

[13] 欧阳宁,包平. 基于本体《中国图书馆分类法》的可视化实现[J]. 图书馆杂志,2008,27(1):28-32.

[14] 田欣. 基于知识本体的图书馆语义检索系统模型研究[J]. 情报杂志,2006,25(6):78-81.

[15] 董慧,杨宁,余传明,等. 基于本体的数字图书馆检索模型研究(Ⅰ)——体系结构解析[J]. 情报学报,2006,25(3):269-275.

[16] 高凡,李景. Ontology及其与分类法、主题法的关系[J]. 图书馆理论与实践,2005(2):44-46.

[17] SCHLICHT A, STUCKENSCHMIDT H. Criteria-based partitioning of large ontologies [C]// Proceedings of the 4th International Conference on Knowledge Capture. New York: ACM press, 2007: 171-172.

[18] STUCKENSCHMIDT H. Network analysis as a basis for partitioning class hierarchies[EB/OL]. [2011-01-12]. http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-171/paper4.pdf.

[19] STUCKENSCHMIDT H, KLEIN M. Structure-based partitioning of large concept hierarchies [C]// Proceedings of the Third International Semantic Web Conference. Hiroshima, Japan: [s.n.], 2004: 289-303.

[20] STUCKENSCHMIDT H, MENKEN M R. Tool support for dependency-based partitioning of OWL ontologies [R], 2005.

[21] KIRYAKOV A, POPOV B, TERZIEV I, *et al.* Semantic annotation, indexing, and retrieval [C]// Proceedings of the 2nd International Semantic Web Conference, LNCS 2870. Berlin: Springer-Verlag, 2003: 484-499.