浙江理工大学学报，第 29 卷，第 4 期，2012 年 7 月
Journal of Zhejiang Sci-Tech University
Vol. 29, No. 4, Jul. 2012

文章编号：1673-3851 (2012) 04-0555-05

语义 Web 环境中基于本体推理的协同标注

祝锡永，周益辉[1]，李　晟[2]

(1. 浙江理工大学经济管理学院，杭州 310018；2. 中国电信浙江省嘉兴分公司，浙江 嘉兴 314033)

摘　要：在已有的 Web 协同标注系统的基础上，通过对资源文档的标签进行共性分析以及上下文情境感知，以此来扩展标签的概念组，并将其与相关本体进行映射；利用本体推理技术来丰富标签的语义性，挖掘出文档隐含的语义信息，发现文档间所存在的内部关联，同时鉴别不同文档之间是否存在着伪关联，以此提高知识检索与知识推荐的准确性以及主体间的知识共享水平。

关键词：语义标注；本体映射；本体推理；SWRL

中图分类号：TP311　　**文献标识码**：A

0　引　言

在 Web2.0 中，用户利用标签对网络中的共享知识资源进行协同标注，使得不同的知识资源拥有不同的描述特征，从而满足知识检索的实现。标注与传统的关键词相比，不但与文档内容相关，而且受用户对于文档的认知程度以及认知的情境等因素的影响，语义性更加发散；同时，不同主体之间的协同标注很好地解决了不同主体可能对于同一资源有着不同的理解，从而影响他们进行抽象标注的问题，提高了对资源描述的全面性以及客观性。但是，这种方法也存在不足之处：a)协同标注系统忽视了标签之间所可能存在的潜在关联，从而否定了所对应资源之间的关联性[1]，这主要表现在同义词上，如标签“student”与“pupil”，两者都有学生的意思，其实际含义相同，可能不同的主体对于描述学生的知识资源采用了这两个不同的标注，但是系统却可能不认定其存在着关联；b)由于计算机理解的局限性，可能莫名增加资源之间的关联性，这主要表现在一词多义上，如标签“mouse”，该词具有老鼠或鼠标两层含义，某一知识资源如果是对于老鼠进行的描述，而另一知识文档是对鼠标进行描述的，两者之间其实不存在联系，但是计算机可能会通过标签“mouse”认为两者之间存在关联。因此，传统的协同标注系统缺乏很好的语义性。

随着语义技术和本体概念的发展，基于本体的语义协同标注正在逐步取代已有的标示技术。从技术角度出发，目前 Web 语义标注研究主要有以下两个方向。

a)基于自然语言的信息抽取方法(natural language processing，NLP)。该方法以标签或者阐述知识资源的自然语句为基础，通过分析句法结构或词性，从句子的主谓宾语法结构中描述出对应的 RDF(resource description framework)陈述，并试图通过 RDF 中的主语和宾语找到对应的领域本体，然后将谓语部分映射为本体属性，从而抽取标签或标注的语义信息[2]。Alani[3] 采用了依赖通用语言本体来完成谓语动词与本体属性关系映射的方法。但是将谓语映射成本体中所对应的属性，存在着一定的困难。同时，这类方法只适用于单个主语或者主谓宾结构的句式，其他句式还不能很好地借用。

b)基于本体的语义标注原型。该方法先对标签进行共性分析，聚类语义相似的标签，然后将其与本体元素直接映射，利用映射所获得的本体信息来

收稿日期：2011－05－05

基金项目：国家自然科学基金资助项目(71071144)

作者简介：祝锡永(1964－)，男，浙江绍兴人，硕士，教授，主要从事信息管理及知识管理的研究。

抽取标签之间的语义关联。文献[4]在 TAP 本体实例集合中查找所有与待标注词匹配的可能实例集合，然后根据待标注词的上下文与实例集合中每个实例的上下文，分别构造各自的文本向量，进行相似度计算，找到与待标注词最匹配的实例。

笔者将从本体推理的角度出发进行研究，基本思路如下：获得用户对于文档的标签，将标签进行格式标准化处理，通过分析标签所处的情境，扩展标签的概念组，并将其与相关本体元素进行映射，接着利用本体的知识结构以及通过本体推理技术丰富标签的语义信息，使得所要处理的标签语义化，从而发现不同文档间所存在的关联或伪关联。

1 标签概念化处理

1.1 标签的格式标准化处理

由于用户对文档所描述的标签，通常可能存在主观的错误，包括分隔符使用错误、拼写错误、使用不兼容符号等错误，由此造成系统不能很好地识别这些标签，所以先要对标签进行格式标准化处理，将其处理成系统能理解和处理的单词格式或简单语句格式。具体分为以下几步：首先，检查标签之间的分隔符是否有使用错误，一般使用“;”号作为分隔符，如果用户使用了“—”或“:”等作为分隔符，由于这些符号可能是某些标签的组成部分，会存在歧义，比如“:”表示时间“14:07”以及“—”表示时间段“20100901—20100930”等，因此要避免使用这几类符号作为分隔符；其次，检查标签的拼写错误，这一步主要检测标签中是否存在错别字或顺序颠倒等问题；再次，检查标签中是否存在英文合成词或缩写词，将合成词拆分成多个词语，将缩写还原成整体词组；最后，获取每个标签的标准表示。

1.2 标签的情景感知处理

情景感知，即上下文关联处理。不同的人可能对于知识资源有着不同的理解，对同一资源所标注的标签也可能不同。单从标签直接出发进行研究，不能很全面地描述资源特征。为此，一方面，分析标签组的共性，即将同一资源的所有标签进行聚类共性分析，即将一个资源的所有标签记为一个初始概念组，统计系统的所有初始组，然后根据初始组间标签的共现，利用阈值找出所有初始组共现最为频繁的标签集，各个标签集即为按含义聚合形成的标签概念组。另一方面，所要做的工作便是自动或人工分析标签在资源文档中所出现的频率，以此来判断标签的精准性。

通过以上两个步骤，可以对标签的概念组进行扩展，以便于使用合适的本体元素来丰富其语义。

2 本体映射以及本体推理技术

获得概念组之后，通过标签与本体之间的映射以及本体与本体之间的映射，为概念组中的每一个概念寻找对应的领域本体。

2.1 本体映射

2.1.1 标签与本体之间的映射(TO-map)

映射的目的是利用语义技术提供的知识结构来丰富标签语义，这样需要将标签的意思中的标签按组与语义网本体的概念、属性和实例映射匹配，从而丰富标签语义并揭示标签间的关系[5]。

可以人为发现标签组对应的领域本体，即通过分析标签概念组中的每一概念，将相关的本体都导入系统中；也可以将概念组的每一个概念通过语义检索系统找出对应的领域本体或者相关语义术语，在这方面做得比较出色的检索系统有 Swoogle[6]。

2.1.2 本体与本体间的映射(OO-map)

本体映射是在两个本体之间存在着一个语义级的关联，能通过这类关联实现本体间的互通性，使得其中一个本体中的实体映射到另一本体的实体上。这里扩展其概念，即定义某一本体是源本体，寻找任一在本体网络中与之存在映射关联的本体，从中寻求更多相关的类、属性或者类的实例。

Ehrig[7]在其文章中给出了一个的本体映射函数：

$map:O_1 \rightarrow O_2$；

如果 $sim(p_1,p_2)>s$，则 $map:O_1 \rightarrow O_2$ 成立，其中 p_1 和 p_2 是两个本体中的实体或概念等；$sim()$表示两个实体间的相似度，根据不同的条件，可以分别对实体的属性、概念名称等进行相似度计算；s 是相似度阈值。

从源本体出发寻找最优关联本体，可以具体划分为以下几个步骤：

a)根据源本体的特征，提取用于计算相似度的特征，如概念、属性名称等；

b)从源本体以及某一相关的本体中选择用于映射的概念或属性对；

c)进行相似度计算；

d)相似度整合：可能从多个指标出发来衡量相似度，得出多种相似度值，因此要对各相似度进行综合统计，从而得到一个整体上的相似度；

e)优化：得到整体相似度之后，这时一般需要人

工的干预,判断映射本体与源本体的关联度,并对其进行调节;

f)迭代第 a)～e)步,直到找到最佳映射关联本体。

通过 TO-map 以及 OO-map,可以找出更多与标签概念组相关的本体、属性、概念以及实体等,这些都能很好地为下一步本体推理工作服务。

2.2　本体推理

找出标签概念组所映射出的对应领域本体,不仅需要这些本体来描述资源,更需要进一步利用这些资源。本体推理的一个基本内容就是由给定的知识获得隐含的知识,在本体中的推理从根本上说就是把隐含在显式定义和声明中的知识通过一种处理机制提取出来。本体的推理有多方面的应用:对于本体的建立者,推理的主要作用是优化表达、检测冲突本体融合;对于本体的使用者,推理的作用主要是获得本体中的知识和运用本体中的知识解决问题。当然,这里所要采用的是后者。

OWL(web ontology language)虽然能很好地描述所处领域本体的抽象概念、实例范例及其相互关系,但对于本体间某些逻辑关系还是不能很好地进行描述,为此在 OWL 基础上引入语义规则语言 SWRL(sematic web rule language)。SWRL 是集本体与规则于一身的一种语言,SWRL 在本体中引入了规则,增强了本体的信息描述能力,提供了更强大的逻辑表达能力[8]。SWRL 的主要目标之一便是提供 OWL 所不支持的表达能力,同时保持与 OWL 语法、语义和理论模型的兼容性,因此可以认为 SWRL 即是一个基于用户自定义规则的补充表达。

SWRL 自定义了一组语法:符号"→"表示蕴涵,它将前提和结论逻辑关联起来;变量以"?"开头;引用的子公式插入符号"^"进行连接;同时提供了许多内置函数,类似于方法调用,返回值为变量的值。

例如:Uncle 问题是 OWL 中经常讨论的一个问题。确定个体 A 是否有个叔叔 U,首先要确定 A 是否有已知的父母 P,其次要确定已知的父母 P 中是否存在男性同胞。如果存在同时满足以上两个条件的数据输出,那么就能确定 U 是 A 的叔叔。

在本体中笔者定义了两个关系:isParent(x_1,x_2)以及 isBrother(x_2,x_3)。isParent(x_1,x_2)表明 x_2 是 x_1 的父母,isBrother(x_2,x_3)表明 x_3 是 x_2 的男性同胞。使用 SWRL 设计一条规则说明 x_1 和 x_3 之间存在有叔叔的关系,如下:

isParent(? x_1,? x_2)^isBrother(? x_2,? x_3)→hasUncle(? x_1,? x_3)

将实体数据 A、U、P 分别代入规则进行检验:isParent(A, P)^isBrother(P,U),这两条规则确实满足,则系统将输出数据 isUncle(A,P),即表明 P 是 A 的叔叔。

上述类型的描述以及关联发现,如果使用普通的标注功能无法实现,必须借助本体的推理功能。对标签进行概念层面上的扩展,形成概念组,找出关联本体及映射本体,按照 SWRL 的描述形式映射为推理规则,并将规则送入推理引擎,启动推理机完成推理。推理完成后,便能得到文档隐含的语义信息。整个过程如图 1 所示。

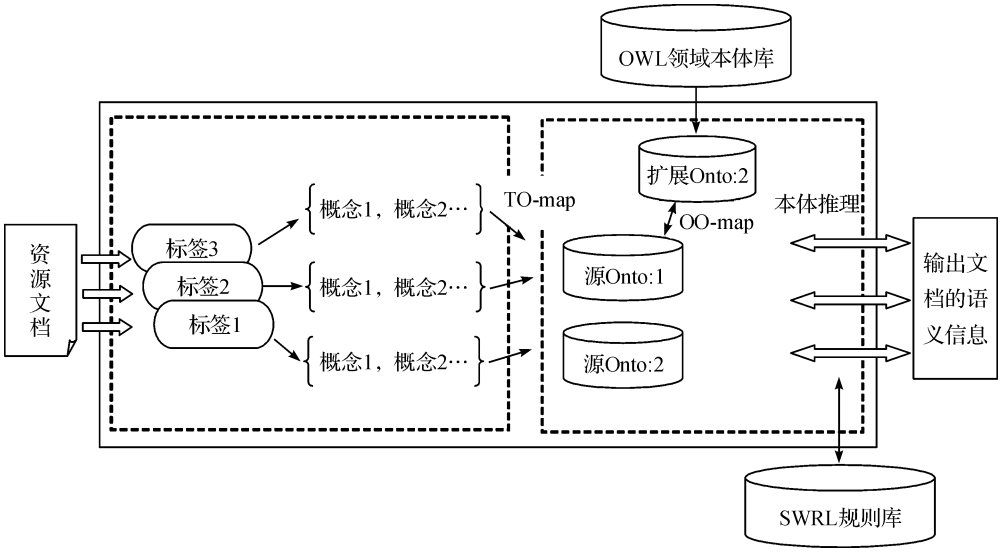

图 1　基于本体推理的标签语义解析过程

3 实 验

下面通过实验来说明以上的理论和方法。设定表 1 所示的两类资源文档。先对其标签进行预处理,即格式标准化处理以及情景感知。

对于文档 A,通过共性分析发现,mouse 与 multi-touch 共现时,就是指 computer-mouse;对于文档 B,根据上下文含义,剔除标签 food,保留 mouse 和 eliminate,且 mouse 与 eliminate 共现时,就是指 mice。通过本体检索系统找到与 mouse 存在映射关系的两类本体(.owl 文件),一类是 mice 本体,一类是 computer-mouse 本体。

表 1 实验资源文档设定

文档	标题	来源	自定义标签
A	Apple:Magic Mouse The world's first Multi-Touch mouse	http://www.apple.com/magicmouse/	mouse; electricity; multi-touch
B	mouse	http://en.wikipedia.org/wiki/Mouse	mouse; eliminate; food

因为构建本体工具 protégé 自身不带推理工具,因此将其与 Jess 推理引擎相结合。将已经得到的两个本体及其相关属性、实例等导入到 Jess 库中,并设定以下 SWRL 推理规则:

Rule－1:hasElectronicProduct(? x_1,? y_1)^hasKey(? x_1,? y_2)^isPartOfComputer(? x_1,? y_3)^hasSignal(? x_1,? y_4)^swrlb:stringEqualIgnoreCase(? x_2,? y_5)

→isComputerMouse(? x_1)

通过以上推理,将标签组的语义信息挖掘出来,发现文档 A 是与 computer-mouse 相关,属于描述电子产品一级,而文档 B 则是关于 mouse(即 mice)的文档,属于描述动物一级。因此,这两类文档间不存在任何关联,并不会因为共有标签 mouse 而存在关联。

进行类似的实验,结果见表 2。通过实验可以发现,笔者所构建的语义协同标注系统丰富了标签的语义信息,能发现不同文档间隐含的关联,也能鉴别不同文档间由于标签含义的多样性而造成的伪关联。

表 2 对比传统标注与语义标注进行文档关联查找的实验结果

文档组	文档标题或 wiki 百科链接	自定义标签	文档之间是否关联	
			传统标注系统	语义标注系统
1	<Protection of Pupil Rights Amendment (PPRA)> <International Student Identity Card (ISIC)>	pupil; PPRA student; ISIC	否	是
2	http://en.wikipedia.org/wiki/Fly <Ready to fly>	fly; fly; song	是	否
3	http://en.wikipedia.org/wiki/Fast_repast http://en.wikipedia.org/wiki/Fast_food	repast; fast food; fast	否	是

建立相对应的本体之后,根据 SWRL 规则进行推导,以上三例的规则建立如下:

1 Rule-1:isPeople(? x_1)^toStudy(? x_1)^hasTeacher(? x_1, True)^swrlb:greaterThan(? age, 8)^sameAs(? x_1, 'pupil')→isStudent(? x_1)

2 Rule-1:isInsect(? x_1)^hasWing(? x_1, True)^hasMobileHead(? x_1)^swrlb:stringEqualIgnoreCase(? x_1)→isTrueFlies(? x_1)

Rule-2:isAction(? x_1)^withWings(? x_1, ? y_1)^isSwift(? x_1)^swrlb:stringEqualIgnoreCase(? x_1,? y_1)^inAir(? x_1)→isFly(? x_1)

3 Rule-1:isKindOfFood(? x_1)^sameAs(? x_1,'repast')^isEatConvenient(? x_1, True)^isEasyToTake(? x_1, True)^swrlb:stringEqualIgnoreCase(? x_1)→isFastFood(? x_1)

4 结 论

笔者提出了利用本体推理技术来挖掘描述资源文档的标签所隐含的语义信息的方法,以此极大地丰富标签的语义信息,帮助系统更好地查找标签所关联的显性或隐含的知识文档,解决传统标注系统中很难解决的同义词、一词多义等问题,提高文档搜索和查找的精确率和查全率,同时也能鉴别不同文档之间是否存在伪关联。这将进一步提高基于 Web 的知识搜索和知识共享水平。下一步工作将是进一步利用不同领域的文体来验证笔者提出的理论和方法。

参考文献:

[1] 吴　芬. 协同标注系统的语义丰富[J]. 情报杂志, 2010, 29(1): 186-188.

[2] 荆　涛, 左万利, 孙吉贵, 等. 中文网页语义标注: 由句子到 RDF 表示[J]. 计算机研究与发展, 2008, 45(7): 1221-1231.

[3] Alani H, Kim S, Millard D, et al. Automatic ontology based knowledge extraction from Web documents[J]. Intelligent Systems, 2003, 18(1): 14-21.

[4] Dill S, Tomlin J. SemTag and seeker: bootstrapping the semantic Web via automated semantic annotation[C]// Proc of the 12th Int'l conf on World Wide Web. New York: ACM, 2003: 178-186.

[5] 高胜利, 施化吉. 语义 Web 中本体推理研究[J]. 淮海工学院学报, 2010, 19(2): 28-32.

[6] Wwoogle[DB/OL]. [2009-09-09]. http://swoogle.umbc.edu.

[7] Ehrig M, Staab S. QOM-quick ontology mapping[C]// Proceedings of the 3rd International Conference on Semantic Web (ISWC). Hiroshima: Springer, 2004: 683-697.

[8] 张艳涛, 陈俊杰, 相　洁. 基于 SWRL 本体推理研究[J]. 微计算机信息, 2010, 26(3):182-184.

Collaborative Annotation Based on Ontology Reasoning in Semantic Web Environment

ZHU Xi-yong[1], *ZHOU Yi-hui*[1], *LI Sheng*[2]

(1. School of Econiomics and Management, Zhejiang Sci-Tech University, Hangzhou 310018, China;
2. Jiaxing Branch Company of Zhejiang Province, China Telecommunications Corporation, Jiaxing 314033, China)

Abstract: Based on Web Collaborative Annotation Systems available, The paper first analyzes the common features of the tags in resource documents and context awareness to extend the concept group of tags and to map with their ontologies, then uses ontology reasoning to enrich the semantic of tags, to mine for implied semantic information, and to describe the internal relationships between documents, finally distinguishes the Pseudo-relevance among different documents in order to improve the accuracy of knowledge retrieval and knowledge recommendation as well as the flow of knowledge among subjects.

Key words: collaborative annotation; ontology mapping; ontology reasoning; semantic web rule language

(责任编辑: 马春晓)