

基于语义关键词的本体特征指数获取方法研究

刘磊, 张波

(同济大学 电信学院 计算机科学与技术系, 上海 201804)

摘要:本体作为知识表达的范例工具,依赖于语义来进行逻辑推理。但在本体搜索时依赖于语义进行搜索十分困难。针对当前本体搜索中存在的问题,提出了一种通过拆分概念来获取语义关键词进而通过计算权值来获得一组本体特征指数的方法。详细阐述了如何获取本体的一组特征指数,即能够描述本体的一组相关的关键词。通过该方法所提炼的语义关键词对于本体搜索以及本体构建工作具有指导意义。

关键词:本体特征指数;描述逻辑;语义网

中图分类号: TP391

文献标识码: A

文章编号: 1673-629X(2008)12-0140-04

Research on How to Get Ontology Index Based on Semantic Keywords

LIU Lei, ZHANG Bo

(Department of Computer Science and Technology, School of Telecommunication,
Tongji University, Shanghai 201804, China)

Abstract: Ontology, as a tool for representation of knowledge, relies on semantics to perform logic inference. However searching an ontology is difficult to manipulate semantics. The method of unfolding concepts of ontology to get a set of keywords with semantics and then computing the weight-value of the keywords to acquire the ontology index is proposed in this paper. How to get the ontology index that a set of keywords which can describe the ontology content and semantic relation is detailed described. The keywords acquired from the methods can instruct the work of ontology searching and constructing.

Key words: ontology index; description logics; semantic Web

0 引言

语义网概念的提出引发了结构化知识模型,特别是本体的研究热情^[1]。本体作为一种知识表示工具,具有良好的概念层次结构和对逻辑推理的支持。在语义网中,本体具有非常重要的地位,是解决语义层次上Web信息共享和交换的基础。

近年来的本体构建热潮使得网上出现了数量庞大的本体库。但随着本体数量的增加,一个新的问题也随之而来:如何更加智能地搜索用户所需要的本体。本体虽然是一种知识表示工具,但它本身却包含了大量的语义信息^[2]。传统的基于关键字的信息搜索方法并不适用于本体搜索。当前的本体构建并没有统一的标准,一般都是领域专家指导本体的构建,这就导致同

一领域内的不同本体库之间会存在很大的差异性。同时本体实例中的某些命名具有一词多义现象,在不同的实例中会有不同的指代含义。这就导致了相同的命名在不同实例中所代表的不同的语义。而且考虑到构建本体时的习惯不同,所使用的概念和属性的命名也会有差异。所有以上情况都使搜索本体工作变得很复杂,仅仅依靠关键字匹配的搜索机制并不能得到很好的结果。

要解决上述问题,必须对已存在的本体进行归纳和总结。本体是大量概念和命名的集合。搜索时不可能将整个本体中的概念集合与输入的查询条件进行直接的匹配。基于以上的限制,提出了本体特征指数这个概念。

本体特征指数必须具备如下两个条件:

- 1) 能够体现原始本体架构的约束规则和关系;
- 2) 能够高度精炼和概括本体内容。

本体特征指数能够反映本体实例内在的语义关系,而且能适用于一些常见的本体操作如搜索、排列等。

文中采用描述逻辑中的公式对本体中的概念进行

收稿日期:2008-03-25

基金项目:国家自然科学基金(70771077)

作者简介:刘磊(1983-),男,硕士研究生,研究方向为语义网、本体协作、电子商务;张波,博士,研究方向为本体论、本体构建与协作、机器学习;导师:向阳,博士,教授,研究方向为智能决策支持系统、本体论与语义网、电子商务。

逐层拆分,提出了个体特征指数的概念以及获得个体特征指数概念的一种方法,为解决个体搜索问题提出了一种新的思路。

1 相关工作

语义网环境下的个体可以以多种形式组织和构建。这就导致了得到个体会有不同的复杂度和组织架构。Yahoo 等门户网站采用了基于分类法来构造一个 Web Directory,得到的是一个比较简单的个体,而著名的 Gene Ontology^[3]则是建立在描述逻辑上的一个结构复杂的大型个体。此外,由于在建模上没有统一的标准,还有许多其他结构各异的个体被创建。在这种情况下,搜索所需要的个体只能依赖于传统的关键词搜索方法再辅以一定的约束条件,通常是比较个体实例的名字和查询条件来获得候选结果如 Swoogle^[4]。这种方法在搜索的准确性和全面性上不尽如人意。为了解决这一问题,很多个体库如 SchemaWeb,采用人工标注的方法来为个体添加适当的标注。还有些个体库通常采用目录结构来存储个体,上传个体的用户来决定自己的个体隶属于哪个目录。仅仅依赖于概念和属性名称来决定个体特征的方法是不合理的,它忽视了个体的结构信息,这对于理解个体所蕴含的语义信息是很重要的。而且由用户来决定个体特征的方法也有问题,没有一种评估机制来确保用户所选择的是正确的目录。

个体特征指数应该是个体中概念的共同特征,并能反映概念间的语义约束关系。我们认为如果能把个体中的概念用一组关键词来表示,那么所有概念所共有的那些关键词就能够体现个体的特征。下面就介绍了如何来获取这组关键词集合。

2 方法设计

个体是对概念明确的规范的说明。个体中的概念和关系必然满足一定的约束条件。个体的语义信息正是通过这些约束条件才得以体现。如果能够得到一组体现该约束条件的关键词,那么这些词就能从一定程度上反映出个体的语义性。称这些词为语义关键词。我们的目标就是从个体中提炼出这些关键词并进而得到能够反映个体本身的特征指数。

文中所提出的方法包含以下几个步骤:

(1)拆分概念。个体中的语义是通过概念间的关系(如 is-a)以及概念的属性来体现的,我们的工作就是将隐藏在概念中的语义使其明晰化。这一明晰化过程可以通过将个体中的概念进行拆分直到元概念(概念和属性均由名称集合中的名称来定义)级别来完成。

个体中的每个概念经过此过程后可以得到与该概念所对应的一组词的集合,其中的每个词都是构成个体的一个元素。这一过程主要依赖于描述逻辑中的产生式变换规则。

(2)计算词的权值,并最终获得个体特征指数。当把所有概念对应的词的集合放在一块时,会发现其中有的词在多个概念中出现,而有些词只在某个概念中出现过一次。被多个概念所涉及到的词,我们认为更能体现个体的内容和特点。通过引入权重计算来获取最终的个体特征指数。如图 1 所示。

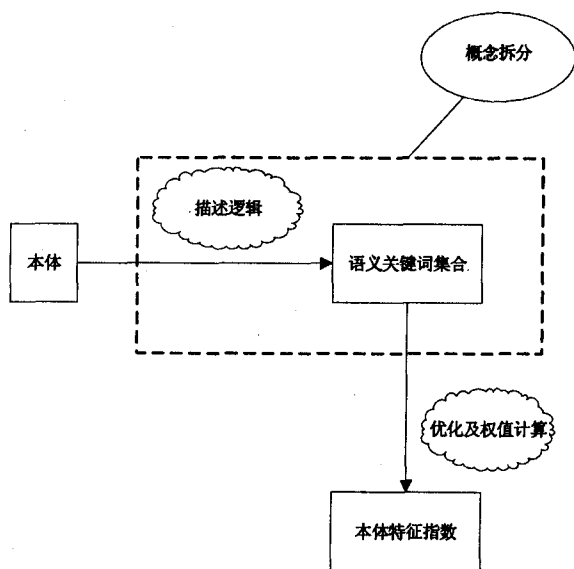


图 1 方法设计思路

2.1 描述逻辑

描述逻辑是一种基于对象的知识表示的形式化,也叫概念表示语言或术语逻辑。描述逻辑(Description Logic)作为一个用于表述以及推理概念知识的逻辑公示集合体,其对语义 Web 的发展起到了非常重要的作用。当前很多个体的构建都是基于描述逻辑的,而且描述逻辑作为个体推理的工具成为很多个体推理语言如 RDF(S)^[2,5],OWL^[5]的原型基础。

通常情况下一个描述逻辑系统包括四个基本组成部分^[6]:表示概念和关系的构造集,TBox 断言集,ABox 断言集,TBox 和 ABox 上的推理机制。

在描述逻辑中,概念被解释为特定领域的子集,关系被解释成该领域上的二元关系,形式上,一个解释 $I = (\Delta^I, \bullet^I)$ 由解释的领域 Δ^I 和解释函数 \bullet^I 构成。解释函数把概念 A 映射到 Δ^I 的子集。描述逻辑的重要特征是其具有很强表达能力的同时又具有可判定性,它能保证推理过程总能停止并返回正确结果。

2.2 获取概念的语义签名

定义个体在描述逻辑中是一个四元组形式 $\langle CS, PS, C, P \rangle$,其中 CS 是概念名称集合,PS 是属

性名称集合, C 是概念集合, P 是属性集合。

描述逻辑中提供了一系列变换规则, 可以用来进行逻辑推理。文中拟采用 DL 产生式变换规则来进行概念的拆分。如图 2 所示。

规则 \cap : $C_1 \cap C_2 \in L(a) \Rightarrow C_1 \in L(a) \text{ 且 } C_2 \in L(a)$

规则 \cup : $C_1 \cup C_2 \in L(a) \Rightarrow C_1 \in L(a) \text{ 或 } C_2 \in L(a)$

规则 \forall : $\forall P. C \in L(a) \text{ 且 } \langle a, b \rangle \in R(P) \Rightarrow C \in L(b)$

规则 \exists : $\exists P. C \in L(a) \Rightarrow \exists b. b \neq a, \langle a, b \rangle \in R(P), C \in L(b)$

规则 \geq : $\geq_n P. C \in L(a) \Rightarrow \exists b_1, \dots, b_k. b_i \neq b_j, \langle a, b_i \rangle \in R(P), C \in L(b_i), k \geq n$

规则 \leq : $\leq_n P. C \in L(a) \Rightarrow \exists b_1, \dots, b_k. b_i \neq b_j, \langle a, b_i \rangle \in R(P), C \in L(b_i), k \leq n$

图 2 DL 产生式变换规则

设 S 为非空实例集合; 任取 $a \in S, L(a)$ 为与 a 相关的概念集合; R 为将属性映射到 $S \times S$ 子集的函数; C 为概念, P 为属性。

通过重复的使用该规则, 可以将一个本体中的一个概念逐层的拆分。采用酒本体^[4]中的一个具体的例子来演示拆分过程。

图 3 为勃艮第红葡萄酒的例子, 下面将列出其拆分过程。对于变换规则的详细描述可参见文献^[7]。

$\text{RedBurgundy} \doteq \text{Burgundy} \cap \text{RedWine}$

$\text{Burgundy} \doteq \text{Wine} \cap \exists \text{ locatedIn. } \{ \text{Bourgogne Region} \}$

$\text{Wine} \sqsubseteq_{=1} \text{hasBody} \cap_{=1} \text{hasColor} \cap_{=1} \text{hasFlavor}$

$\cap_{=1} \text{hasMaker} \cap_{=1} \text{hasSugar} \cap \forall \text{ hasMaker. Winery}$

$\cap \exists \text{ locatedIn. Region} \cap \geq_1 \text{madeFromGrape}$

$\text{RedWine} \sqsubseteq \text{Wine} \cap \exists \text{ hasColor. } \{ \text{Red} \}$

图 3 RedBurgundy 示例

${}^0S^{\text{RB}} = \{x: \text{Burgundy} \cap \text{RedWine}\}$

${}^1S^{\text{RB}} = \{x: \text{Wine} \cap \exists \text{ locatedIn. } \{ \text{BourgogneRegion} \} \cap \exists \text{ hasColor. } \{ \text{Red} \}$

${}^2S^{\text{RB}} = \{x: (\cap_{=1} \text{hasBody} \cap_{=1} \text{hasColor} \cap_{=1} \text{hasFlavor} \cap_{=1} \text{hasMaker}$

$\cap_{=1} \text{hasMaker} \cap_{=1} \text{hasSugar} \cap \forall \text{ hasMaker. Winery}$

$\cap \exists \text{ locatedIn. Region} \cap \geq_1 \text{madeFromGrape})$

$\cap \exists \text{ locatedIn. } \{ \text{BourgogneRegion} \} \cap \exists \text{ hasColor. } \{ \text{Red} \}$

$\cap \text{RedWine} - \text{spec} \cap \text{Wine} - \text{spec}\}$

${}^3S^{\text{RB}} = \{ \langle x, y0 \rangle : \text{hasBody}, \langle x, y1 \rangle : \text{hasColor}, \langle x, y2 \rangle : \text{hasFlavor},$

$\langle x, y3 \rangle : \text{hasMaker}, \langle x, y4 \rangle : \text{hasSugar}, y3: \text{Winery},$

$\langle x, y5 \rangle : \text{locatedIn}, y5: \text{Region}, \langle x, y6 \rangle : \text{madeFromGrape},$

$y1: \{ \text{Red} \}, y5: \{ \text{BourgogneRegion}, x: \text{Wine} - \text{spec} \} \}$

${}^4S^{\text{RB}} = \{ \langle x, y0 \rangle : \text{hasBody}, \langle x, y1 \rangle : \text{hasColor}, \langle x, y2 \rangle : \text{hasFlavor},$

$\langle x, y3 \rangle : \text{hasMaker}, \langle x, y4 \rangle : \text{hasSugar}, y3: \text{Winery},$

$x: \text{Wine} - \text{spec}, \langle x, y5 \rangle : \text{locatedIn}, y5: \text{BorgogneRegion} - \text{spec},$

$\langle x, y6 \rangle : \text{madeFromGrape}, y1: \text{Red} - \text{spec}\}$

当满足 ${}^nS = {}^{n-1}S$ 时, 拆分过程结束。可以证明, 对于没有循环结构的本体, 在选定合适的产生式的情况下, 该拆分过程是可以终止的^[8]。当前很多 DL 系统中对产生式变换规则已经实现并且优化, 在处理实际本体时的性能是可以接受的。

通过拆分概念, 隐含在本体中的语义信息可以显式地表现出来。本体是对概念的明确的规范化说明, 那么本体中的每个实例必然要满足该约束。通过拆分概念, 能够知道本体中的实例是如何满足该规范并且获取该约束条件。最终得到的结果是本体中的每个概念都会有一组由元概念和元属性组成的标记集合与之对应。可以证明由该标记集合可以反向生成原始概念。

2.3 获取本体特征指数

由于建模时存在的差异性如拼写的差异等, 所得到的标记集合需要进行优化, 以消除差异性。

借助自然语言处理技术, 可以将标记集合中的关键词进行处理, 如对于属性 hasXXX , 将其处理为 XXX , 因为 has 并不包含太多所需要的语义信息。其次需要把所有的词统一为小写形式。

本体所包含的概念数目是有限的, 因此通过拆分所得到的标记集合也是有限的。通过优化得到的标记集合, 与本体中的每个概念都是相关的, 因此它们可以作为本体特征指数的候选词。

每一个语义标记词都对本体的某个方面进行了说明。例如在酒本体中, winery 是表示酿造酒的工厂, 而 color 表示酒的颜色。而且在概念的拆分过程中, 某个语义标记词可能会被多个概念所涉及。我们认为, 如果某个语义标记词被本体中绝大多数概念都涉及到, 那么该标记词在本体中是比较重要的。为了体现这种思想, 为标记词引入了权值。通过计算每个标记词的权值, 最后权值最大的 n 个标记词作为最终的本体特征指数 (n 可以人工指定)。

权值的计算公式如下:

* 以 $\langle x, y \rangle$ 或 x 开头的标记词的初始权值定为 $A_0 = 1$;

* 由属性 P 引入的标记词的权值定为 $A = A_p * A_0$;

* 若某标记词 e 在 n 个概念的拆分结果中出现则置其权值 $A_e = \sum_{i=1}^n A_i^e$;

设 s 为一个标记词, s 的最终权值计算公式为:

$$W_s = \frac{A_s}{\sum_{e \in O} W_e}$$

其中分母为所有语义标记的权值之和。最终计算结果

中权值较大的关键词即可作为本体的特征指数。最终得到的本体特征指数形式如下所示:

$$\{Item1/W_{ime1}, Item2/W_{ime2}, \dots, Itemn/W_{imen}\}$$

得到了这组特征指数后,可用来对本体进行标注。在本体搜索时就可根据输入的查询条件与特征指数进行匹配,并按照权值的大小对结果进行排序。

3 结束语

本体是语义 Web 中各种智能应用的基本工具。在当前本体数量越来越多且又没有统一的本体建模标准的情况下,高效的检索本体变的越来越重要。介绍了一种基于描述逻辑产生式算法的通过语义关键词来标记本体的方法。文中的主要贡献在于通过对本体所包含的概念进行拆分和标记,使得本体与一组能够高度概括其语义信息的关键词联系起来,通过计算各关键词的权值,选出本体中最能体现本体特点的一组关键词,这组关键词就成为该本体的特征指数。

利用描述逻辑中 Tableaux 算法进行概念拆分的过程是比较耗时的,它的时间复杂度最坏的情况下是多项式级别的。但这并不影响该方法的可用性。

该方法还有待解决的几个问题:

1)对甄选结果的评估。当前的评估只能借助于领域专家来评判。自动评估机制有待进一步的研究;

2)如何通过拆分出来的关键词反向生成本体。当

前的本体构建主要是手工或半自动构建。文中提出的方法如何用来指导本体构建和设计也是下一步研究工作的重点。

参考文献:

- [1] Berners-Lee T, Hendler J, Lassila O. The Semantic Web [J]. Scientific American, 2001, 284(5): 28-37.
- [2] Dean M, Schreiber G, van Harmelen F, et al. OWL Web Ontology Language Reference [EB/OL]. 2003-08-18. <http://www.w3.org/TR/owl-ref/>.
- [3] Hu B, Dasmahapatra S, Dupplaw D, et al. Reflections on a medical ontology [J]. International Journal of Human-Computer Studies, 2007, 65(7): 23-25.
- [4] Ding L, Pan R, Finin T, et al. Search on the semantic Web [J]. IEEE Computer, 2005, 10(38): 62-69.
- [5] RDF Core Working Group, RDF Reference [EB/OL]. 2004. <http://www.w3.org/RDF>.
- [6] 史忠植,董明楷,蒋运承,等.语义 Web 的逻辑基础 [J]. 中国科学 E 辑. 信息科学, 2004, 34(10): 1123-1138.
- [7] Baader F, Calvanese D, McGuinness D, et al. The Description Logic Handbook: Theory, Implementation and Applications [M]. Cambridge: Cambridge University Press, 2003.
- [8] Horrocks I, Sattler U. A tableaux decision procedure for SHOIQ [C] // In Proc. of the 19th Int. Joint Conf. on Artificial Intelligence (IJCAI 2005). [s. l.]: Morgan Kaufman, 2005.

(上接第 139 页)

围,增强服务效能。

基于 Web Services 异构资源汇集与共享系统很好地实现了该项目的宗旨,通过客户端软件将异构资源进行标准化处理,在服务器端获取数据资源和配置信息进而存储,屏蔽了资源的异构性;用户通过网络科技资源应用集成环境平台门户网站检索需要的资源,使用户对资源的访问透明化。

4 结束语

文中是在参与国家科技基础条件平台应用服务支撑系统的子项目“网络科技资源应用集成环境平台建设”的过程中,针对网络科技资源分布广泛、资源共享程度不高、资源信息形式多样、种类繁多等特点,利用 Web Services 技术,按照面向对象的设计方法,设计并实现了基于 Web Services 异构资源汇集与共享系统。该系统能有效地屏蔽科技资源各领域平台的差异,实现统一的资源传输规范,使国家各领域平台的科技资源能够快速、准确、有效地被整合利用,目前该系统已上线试运行。

参考文献:

- [1] 马大川,杨红平.信息资源的集成整合研究 [J]. 中国图书馆学报, 2004, 30(3): 36-40.
- [2] Li Jun, Furuse K, Yamaguchi K. Focused Crawling by Exploiting Anchor Text Using Decision Tree [C] // Proceedings of the 14th International World Wide Web Conference. Chiba, Japan: [s. n.], 2005: 1190-1191.
- [3] Cheng Jing, Li Qing, Wang Li ping, et al. Automatically generating an e-textbook on the Web [C] // In: Lecture Notes in Computer Science. Berlin: Springer-Verlag, 2004: 35-42.
- [4] 柴晓璐,梁宇奇. Web Services 技术、架构和应用 [M]. 北京:电子工业出版社, 2003.
- [5] Deitel H M, Deitel P J, Duwaldt B, et al. web 服务实用技术教程 [M]. 北京:机械工业出版社, 2004.
- [6] 屈良.基于 Web Services 的网络信息资源集成研究 [J]. 中国信息导报, 2006(9): 56-61.
- [7] 林彤,舒真才.基于 web Services 的异地异构数据库的集成 [J]. 北京工业大学学报, 2005, 31(2): 210-213.
- [8] 吴建中. DC 元数据 [M]. 上海:上海科学技术文献出版社, 2000.