面向Web信息资源的领域本体模型自动构建机制的研究

金 鑫
(中央财经大学信息学院 北京 100081)

摘 要 领域本体的构建是本体工程研究与应用的重要内容。面向网络Web信息资源,获取领域相关文本信息,通过对文本的概念分析,构建领域本体模型。提出一套本体自动构建机制,该本体构建基于数据挖掘和机器学习技术,内容主要包括基于贝叶斯(Bayes)分类原理;提出多个分类器方式的概念分类过程和算法;提出概念关联分析和概念自学习算法,建立本体原型;提出面向OWL本体模型的转换映射机制,构建基于OWL的本体模型。此外,还提出了从网络资源获取、领域本体建模到本体实施应用的一套完整的本体构建和应用实施的解决方案。

关键词 本体,概念分析,本体自动构建,OWL
中图法分类号 TP393 **文献标识码** A

Research on Mechanism of Automatic Construction of Ontologies for Web Information Resources

JIN Xin
(School of Information, Central University of Finance and Economics, Beijing 100081, China)

Abstract To build domain ontology is an important part of ontology research and application areas. This paper proposed the process and algorithm of automatic construction of ontologies with data mining and machine learning technologies for web information resources. During the procedure of ontology construction, this paper firstly presented the process and algorithm of concepts classification with muti-classifiers based on Bayes classification principle, then discussed the algorithms of concepts-associated analysis and concept self-learning to build the ontology prototype, whereafter presented ontology transforming mechanism from ontology prototype to OWL ontology. This paper also proposed a whole system solution from mining web pages, building domain ontology to the end of ontology application.

Keywords Ontology, Concept analysis, Automatic construction of ontology, OWL

1 引言

从20世纪90年代,关于本体(Ontology)工程的研究,在信息科学领域开始受到关注[1]。近年来,本体已经成为知识分类、知识表达、知识共享和重用等相关研究的重要手段,及知识建模、语义集成与互操作处理的核心环节。构建本体是应用本体的前提,Uschold最早提出了一个本体构造的方法指导框架[2],该指导框架为本体构建研究提供了一般性的指导原则。原来的手工本体构建方法工作量大、繁琐、效率低,已经成为本体设计开发的主要瓶颈。因此,研究人员越来越关注本体自动化或半自动化构建研究。

目前关于本体自动化或半自动化构建的方法论的研究成果较少。Maedche等提出了需要人工参与的半自动化本体学习框架[3],采用稳定协作模型的范例用于构建语义网上的本体,这个框架使用半自动化的本体构建工具扩展了传统的本体工作环境。Zhong等提出了面向专门领域本体构建过程的方法[4],在这个分阶段的本体构建过程中,使用了各种文本挖掘技术和自然语言处理方法[5]。概念的分析是本体构建的重要内容,文献[6]通过分析概念网络关系来研究本体模型的自动化构建,但是其本体自动化构建的前提假设是概念网络已知;文献[7]通过形式化概念分析研究本体构建方法的优化。此外,目前较为流行的是基于通用词典(如WordNet和HowNet)和自然语言处理的本体半自动构建的研究[8,9],但这种方式构建领域本体的前提是预先定义足够丰富和清晰的领域术语集分类以及概念之间的关系,然后才能通过概念学习来发现相关的概念知识,也就是说前提是假定本体元模式已知;近似的研究还有在已知关系数据库模式的基础上自动转换生成本体,如文献[10]。

本研究与现有其他研究的不同之处主要体现在3个方面:其一,数据来源上,本研究数据来源于海量的网络空间(Web页面),面向的应用领域结合依托的相关项目主要面向纺织服装领域;其二,研究方法上,本研究并未假定清晰已知的领域概念空间,而是通过机器学习的方式,基于多分类器进行概念自动分类,并进行概念的自动学习,所用的术语字典,只是描述常用的领域术语以及相关的同义词、近义词,并不作为最终的概念分类标签;其三,完整的研究机制和实施方案。本研究从最初的网络文本获取、文本的概念分析、本体原型的构建和求精、面向OWL[11](Web Ontology Language,网络本

到稿日期:2011-08-24 返修日期:2011-11-23 本文受国家自然科学基金(61100112),中央财经大学党建课题(DJD11006),中央财经大学重点学科资助。

金 鑫(1974—),男,副教授,主要研究方向为知识工程、商务智能,E-mail:james.jin2009@gmail.com。

体语义)本体模型的映射、OWL 的本体知识表示,一直到如何实际应用本体模型,提供了一套完整的研究机制和实施方案。

在本文的研究中,面向 Web 网络资源,使用网络爬虫爬取文本信息,然后提出一套本体自动构建机制,用来分析领域文本概念,构建并表示本体模型。本体构建过程中,基于贝叶斯(Bayes)分类原理,提出多个分类器方式的概念分类过程和算法,并对其进行概念分类;提出概念关联分析和概念自学习算法,生成本体原型;建立面向 OWL 本体模型的映射和表示机制,生成最终的 OWL 本体模型。最后对本体的应用和实施进行了阐述。

2 系统架构

本研究设计了以本体自动构建为核心的 DOBS 系统(Domain Ontology Building Service System),图 1 是 DOBS 系统架构图。该系统首先通过网络爬虫 Crawler 工具从指定网站获取大量文本信息,并以静态网页文本的方式存储在本地。本体的构建从对文本的分类标注开始,依次进行概念的关联分析和自学习、本体原型的生成求解、OWL 本体模型的映射和知识表示,从而提出一套本体自动构建机制。

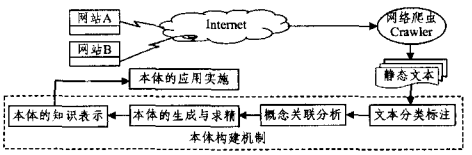

图 1 DOBS 本体自动构建系统架构

DOBS 系统中的网站爬虫 Crawler 是一个站点页面爬取和结构分析工具。它从给定的站点起始页面开始,可以按照广度优先或深度优先的方式遍历站点,保存页面到本地、修改页面的文件名、修改页面中的链接(这与离线阅览器的功能类似)。它能产生站点结构的摘要,包括页面模板的数量和接口信息。

本文第 3 节将介绍 DOBS 系统中本体自动构建的过程和算法,第 4 节将介绍本体的应用实施。

3 本体自动构建机制

本研究提出的自动化本体构建机制主要包括以下过程:1)文本分类标注;2)概念关联分析;3)本体原型的生成和求精;4)面向 OWL 本体模型转换和 OWL 本体知识表示。在整个过程中使用了领域术语字典作为整个背景知识的语境基础,并且整个过程是可循环递归求解的。以下各小节将阐述本体自动化构建过程的机制和算法描述及相应的结果。

3.1 文本分类标注

为了从各种形态的文本资源中挖掘概念,构建本体,首先需要用分类的机制标注各种文本资源。如果用手工的方式标识分类文本资源,工作量太大。因此,这里基于非确定取样和随机分类器,提出了半自动化的方式以研究文本资源的分类。

本研究借鉴了 Lewis 和 Zhong 等关于分类器的非确定取样算法[4,12],基于贝叶斯分类器原理,使用多分类器方法进行多种类的分类。

根据贝叶斯公式[13],一个分类器可以估计后验概率:

$$P(C_i|w)=\frac{P(w|C_i)\times P(C_I)}{\sum_{j=1}^{q}P(w|C_j)\times P(C_j)} \tag{1}$$

式中,C_i 是某样本归属的一个不相交穷举类集;$w=\{w_1,w_2,\cdots,w_d\}$,是经过直觉概念形态分析的概念术语集,是一个观测向量,$P(w|C_i)$是归属类 C_i 的样本有观测向量 w 的条件概率。$P(C_i)$是样本归属类 C_i 的先验概率。$P(C_i|w)$是通过观测向量 w 获得的向量 C_i 的估计。

如果只假定有两种分类情景,即 $q=2$,则只有两种类别 $C_1=C,C_2=\overline{C}$,且 $P(\overline{C})=1-P(C)$。C 和 $\overline{C}$ 的相对后验概率可以用式(2)的后验条件几率描述:

$$\frac{P(C|w)}{P(\overline{C}|w)}=\frac{P(C)}{P(\overline{C})}\times\frac{P(w|C)}{P(w|\overline{C})} \tag{2}$$

对于给定的大量观测集 w,其条件几率 $P(w|C)/P(w|\overline{C})$通过直接观察估计,可以对式(2)有以下的分解:

$$\frac{P(C|w)}{P(\overline{C}|w)}=\frac{P(C)}{P(\overline{C})}\times\prod_{i=1}^{d}\frac{P(w_i|C)}{P(w_i|\overline{C})} \tag{3}$$

因为 $P(\overline{C}|w)=1-P(C|w)$,经算术运算处理,有:

$$P(C|w)=\frac{\exp(\log\frac{P(C)}{1-P(C)}+\sum_{i=1}^{d}\log(P(w_i|C)/P(w_i|\overline{C})))}{1+\exp(\log\frac{P(C)}{1-P(C)}+\sum_{i=1}^{d}\log(P(w_i|C)/P(w_i|\overline{C})))} \tag{4}$$

式中,虽然只分成了两个类 $C_1=C,C_2=\overline{C}$,且 $P(\overline{C})=1-P(C)$,但是可以通过对式(4)进行扩展来处理多个类的分类。

基于上面的阐述,概念文本分类算法的主要过程如下:

Step1 用户首先选择一系列术语样本作为初始分类器分成 N 类,所有 N 个类初始认定为组成一个反面类集合。

Step2 从反面类集中选择一个类作为正面类,而剩余的其它类仍然构成一个反面类集。

Step3 用户分类概念文本。

Step3.1 对每个未分类的文本使用目前的分类器分类。

Step3.2 找到 k 个文本,通过使用式(4)计算后验概率至少可以确定类成员。

Step3.3 用户分类 k 个文本的子样例。

Step3.4 对所有已分类的概念文本使用一个新的分类器训练。

Step4 重复执行 Step2,Step3 直到所有的分类都被选作一个正面类。

对于专门应用的需要,用户选择样例作为初始的分类器是很重要的一步,在分类器中表示了用户的需要和用途。在本研究中以纺织服装领域的术语集作为初始的分类器样本。

3.2 本体原型的生成

基于上面的概念文本分类结果,本体生成过程包含两个阶段。

第一个阶段是概念化关系分析(conceptual relationship analysis)。首先通过式(5)和式(6)各自计算文本中概念术语的组合权重。

$$D_i=\log d_i\times tf_i \tag{5}$$

$$D_{i,j}=\log d_{i,j}\times tf_{i,j} \tag{6}$$

式中,d_i 和 d_{ij} 是文本频度,d_i 表示在 n 个文本组成的集合中,有术语 i 出现的文本数量。d_{ij} 表示在 n 个文本组成的集合中,有术语 i 和术语 j 同时出现的文本数量。tf_i 和 tf_{ij} 是术语频度,tf_i 表示术语 i 在文本中出现的次数,tf_{ij} 表示术语 i 和术语 j 在文本中同时出现的次数。

可以通过式(7)和式(8)计算术语之间的近似关系生成网络化的概念空间。

$$Rel(i,j)=\frac{D_{i,j}}{D_i} \tag{7}$$

$$Rel(j,i)=\frac{D_{i,j}}{D_j} \tag{8}$$

式(7)计算术语 i 到术语 j 的关联，式(8)计算术语 j 到术语 i 的关联。我们也使用阈值以确保只有最相关的术语被保留。以上一节介绍概念文本处理后的服装销售(garment marketing)概念文本集为例，对其进行内容术语检索，通过本节的上述公式计算概念术语的关联，由于其程序自动运算过程简单，这里省略其算法。表 1 描述了经程序自动运算后得到的服装销售中部分概念术语近似关联。

表 1　术语的近似关联

Term i	Term j	Rel(i,j)
款式	服装	0.8389
面料	服装	0.7341
类型	服装	0.5409
设计师	服装	0.4929
企业	服装	0.4636
中国	服装	0.4033
市场	服装	0.1903
品牌	服装	0.4682
面料	款式	0.8643
服装	款式	0.7438
类型	款式	0.4059
企业	款式	0.2764
设计师	款式	0.3891
中国	款式	0.2854
品牌	款式	0.5433
…	…	…

第二阶段使用 Hopfield 网络变量生成本体原型。Hopfield 网络中的节点值基于一个局部计算原则被迭代更新，这个局部原则是：每个节点的新的状态仅依赖于在某一时刻输入的网络权值。我们把每个剩余术语称作一个神经单元，认定术语 i 和术语 j 的近似关联是单向的，单元之间的连接被加权。在 0 时刻有：$\mu_i(0)=x_i$，$0\leqslant i\leqslant n-1$，其中 $\mu_i(t)$ 是单元 i 在 t 时刻的输出，x_i 表明了单元的输入模式，其值为 0 或 1。在时刻 0 只能有一个概念术语接收到值 1，而其它术语的接收为 0。假定只有 n 个概念术语，我们可以重复 n 次式(9)：

$$\mu_j(t+1)=f_s[\sum_{i=0}^{n-1}w_{ij}\mu_i(t)],0\leqslant j\leqslant n-1 \tag{9}$$

式中，w_{ij} 表示了近似关联 $Rel(i,j)$，f_s 是一个反曲函数，如式(10)：

$$f_s(net_j)=\frac{1}{1+\exp[(\theta_j-net_j)/\theta_0]} \tag{10}$$

式中，$net_j=\sum_{i=0}^{n-1}w_{ij}\mu_i(t)$，$\theta_j$ 是一个阈值，θ_0 用于调整反曲函数的外形。

重复这一过程，直到相邻两次的迭代输出的术语不再改变为止，可以通过式(11)检查迭代过程是否会聚。

$$\sum[\mu_j(t+1)-\mu_j(t)]^2\leqslant\varepsilon \tag{11}$$

式中，ε 是允许的最大错误率。

最后的输出表现为与最初术语相关的概念术语集合，这可以认为是本体的原型。图 2 描述了一个关于服装销售(garment marketing)的本体原型，其中粗线表示了概念之间是强关联，细线表示了概念之间是弱关联，该本体原型是以表 1 中的每个术语为初始输入模式，经 Hopfield 网络学习得到的结果。

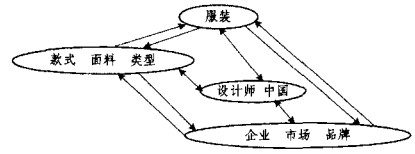

图 2　一个关于服装销售的本体原型片断

3.3　本体的求精

在本体的构建过程中，如果尽可能合并任何关联的知识，有助于提升整个过程的功效和本体生成的质量。而存储背景知识的字典就是一个有用的资源，它可以作为本体求精的背景知识，通过使用字典，术语可以被它们的同义字及其广义和狭义含义扩展。图 3 使用了字典的关于服装销售(garment marketing)的本体，相比不使用字典的本体构建，它的本体质量更高，效果更好。注意图 3 所示的概念空间与图 2 所示相同，图中示出了 A、B、C、D 这 4 个概念空间。

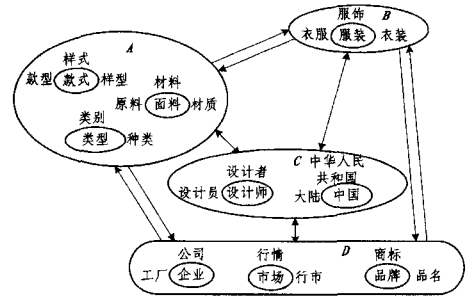

图 3　使用字典的本体求精

3.4　本体的表示

基于上述的本体构建机制建立了领域本体，其实质是对纷繁芜杂的信息源进行概念分析和概念挖掘，并发现概念分类和概念关联，建立本体原型。基于本体原型，需要使用本体表示语言来表示本体。本项目中采用流行的 OWL 本体表示语言对本体进行描述。

OWL(Web Ontology Language)是目前最为流行的本体知识表示语言，它已经成为 W3C 推荐的本体标准规范。OWL 具有本体表示语言所需的全面的知识表达能力及良好的语法和语义互操作性，而且它还是基于 XML 标准规范的，并且兼容对 XML 和 RDF(S)语法和语义的扩展支持。

本研究将前述建立的本体原型转换为基于 OWL 的本体模型，其主要进行两部分的工作，其一，本体原型向 OWL 本体模型的映射，其二，基于 OWL 的本体表示。

OWL 表示的本体模型的内容主要包含类(classes)、属性(Properties)和个体(Individuals)。其中属性分为描述静态特征的数据类型(datatype Properties)和类对象之间关系的对象属性(Object Properties)，因为本体的 Individuals 内容可以和具体的项目实施关联，将放在下一小节中介绍。这里关于本体原型向 OWL 本体映射的机制，就是要解决如何将原型中的概念以及概念之间的关系向 OWL 的内容进行转换。

从本体原型向 OWL 本体模型转换的过程如下：

Step1　从某一概念空间中抽取核心概念，将核心概念定义为一个 Class，例如从概念空间 D 中选择概念企业(Enterprises)作为核心概念，即生成一个主概念类 Enterprises。

Step2　判断该概念空间中的其他概念是否具有子属性，是否和其它类对象关联，如果该概念不再具有子属性(不含诸如 Name 这样的固有属性)，而且不和其它外部对象发生关

联，则将该概念作为主概念的数据类型属性；如果该概念具有子属性，则将该概念转换为主概念类的子类。例如概念空间 D 中，市场情形描述(market_description)作为企业的一个数据对象属性，品牌由于和服装对象相关联，则其经该概念转换为主概念类的一个子类 Brands。

Step3　根据前述两步骤，依次转换映射每个概念空间的内容。

Step4　对于难以确定核心概念的概念空间，且其中的每个概念不具有子属性，则该概念空间中的概念确定为其他概念空间的属性。例如概念空间 A 中的款式、面料、类型等概念分别映射为服装类 Garments 的数据属性 Style、Material、Type。

Step5　对于发生关联的概念空间，描述类之间的关系，即定义对象属性。

Step6　手工完善 OWL 本体模型，主要包括补充通常的数据属性，如 NAME、ID、备注等；完善对象之间的关系和限定，如增加对属性区域(Domain)和范围(Range)的限定等。

根据上述步骤，可以将前述如图 3 所示构建的本体原型转换为 OWL 表示的本体模型，如图 4 所示。图 4 中的左侧部分为 OWL 图形化表示，右侧为 OWL 本体的知识描述。

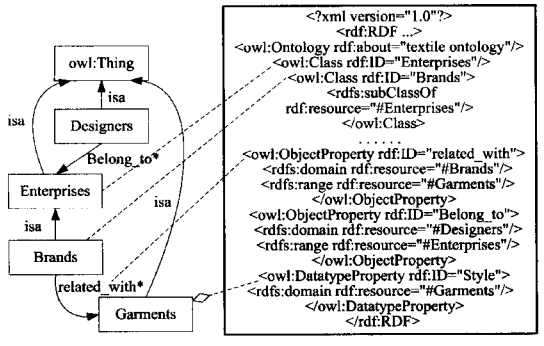

图 4　OWL 本体模型及其知识表示片段

4　本体的实施应用

前述构建了 OWL 本体，该本体模型以 OWL 文件的形式存在，但该本体与实际应用结合还缺少两部分内容：实例信息(即 Individuals 个体)和知识规则(如知识检索、知识推理)。关于实例信息，可以 Individuals 个体的方式成为 OWL 本体模型内在的一部分，也可以采用 OWL＋关系数据库(MySql)的方式，将本体的知识模式(Schema)与具体的数据内容分开，即实例数据存储在关系数据库中。因为 OWL 和 Mysql 可以很容易进行关联集成，而且考虑到实例信息的海量化、未来的扩展应用(如基于数据库的查询、整合等)，所以本研究项目采用了 OWL＋Mysql 的方式，数据实例信息以记录方式存储在关系数据库表中。

基于本体的知识规则可以内置在本体中(如定义 swrl 规则)，也可以通过外部的 OWL 的推理和查询引擎来实现(如 Jena＋Jess＋Sparsql)。本项目研究中由于数据内容存储在数据库中，因此主要通过外置的本体推理和查询引擎来实现知识的查询和推理。

为了实现前述的研究，本项目构建了 DOBS 系统平台，图 5 所示的是该系统的某个工作界面。该系统平台中集成了对网络信息的爬取、本体的构建、本体内容的展示等功能。在对本体构建和知识表示的过程中，本项目同时应用了 Protégé 平台提供的功能。

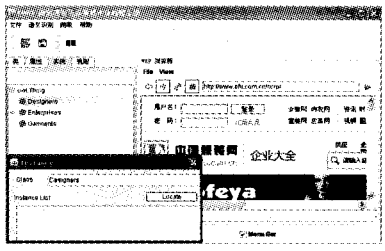

图 5　DOBS 系统平台 GUI 界面

结束语　对本体构建的研究，一直是本体工程研究的热点，由于手工构建方法效率低、出错率高，因此，本文研究中引入了知识发现和人工智能技术，提出了本体自动构建方法。文中面向 Web 网络资源，使用网络爬虫爬取文本信息，然后针对静态文本进行本体自动构建。本体构建过程中，首先基于贝叶斯(Bayes)分类原理，扩展了单个分类器方式，提出多个分类器方式的概念分类过程和算法；然后通过概念关联分析和 Hopefield 原理，提出概念自学习算法来构建本体原型；接着探讨了本体原型到 OWL 本体模型的映射机制。最终形成了一个完整的本体自动化构建模式。本文的研究讨论了从 Web 文本信息的获取到最终的本体应用实施，探索形成了一套完整的本体构建和应用实施的系统解决方案，相信会对相关的研究人员有一定的借鉴作用。

参 考 文 献

[1] Gruber T. Toward principles for the design of ontologies used for knowledge sharing[J]. International Journal Human Computer Studies, 1995, 9: 907-928

[2] Uschold M, Tate A. Putting ontologies to use [J]. The Knowledge Engineering Review, 1998(13): 1-3

[3] Maeduche A, Staab S. Ontology Learning for the Semantic Web [J]. IEEE Intelligent Systems, 2011, 16(2): 72-79

[4] Zhong N, Yao Y Y, Kakenoto Y. Automatic Construction of Ontology from Text Databases[J]. Data Mining, 2011, 2: 173-180

[5] Tao D, Li Yue-feng, Zhong Ning. A Personalized Ontology Model for Web Information Gathering [J]. IEEE Transactions on Knowledge & Data Engineering, 2011, 23(4): 496-511

[6] 吴江. 互联网资源知识本体自动构建实证研究[J]. 图书情报工作, 2011, 55(11): 116-120

[7] 张云中，徐宝祥. 基于形式概念分析的领域本体构建方法优化研究[J]. 图书情报工作, 2010, 54(8): 112-115

[8] 徐力斌，刘宗田，等. 基于 WordNet 和自然语言处理技术的半自动领域本体构建[J]. 计算机科学, 2007, 34(6): 219-222

[9] 丁晟春，傅柱. 基于航天叙词表的领域本体半自动化构建研究[J]. 情报理论与实践, 2011, 34: 113-116

[10] 吕艳辉，马宗民，王玉喜. 基于关系数据库的 OWL 本体构建方法的研究[J]. 计算机科学, 2009, 36: 153-156

[11] OWL Web Ontology Language[EB/OL]. http://www.w3.org/TR/owl-features/, 2004

[12] Lewis D D, Catlett J. Heterogeneous Uncertainty Sampling for Supervised Learning[C]// Proc. 11th Inter. Conf. on Machine Learning. 1994: 148-156

[13] Liu Bing. Web 数据挖掘[M]. 余勇，等译. 北京：清华大学出版社, 2009

[14] 张鹏，王国胤，陶春梅，等. 基于本体粗糙集的程序代码相似度度量方法[J]. 重庆邮电大学学报：自然科学版, 2008, 20(6): 737-741