面向Web结构化信息处理的汉语知识库构建研究

郭文宏 范学峰
(同济大学电子与信息工程学院 上海 201804)

摘 要 对Web结构化汉语信息处理中的知识需求进行了分析,介绍了目前有影响的汉语语义资源和本体知识,给出了面向Web结构化信息处理的汉语知识库组成模型及构建方法,并在Deep Web研究中对该模型进行了应用验证。该研究旨在使计算机更全面有效地对特定领域的Web结构化信息进行处理,对本体的深入研究也有一定的参考价值。

关键词 Web结构化信息,信息处理,知识库构建,语义资源,本体知识

On Knowledge Base Construction for Structured Chinese Web Information Processing

GUO Wen-hong FAN Xue-feng
(College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China)

Abstract This paper analysed the requirements of knowledge for the structured information processing on Web and introduced the Chinese semantic resources and ontology nowadays. A knowledge base model was proposed for Chinese information processing on Web. A method to construct was given. The model was applied in Deep Web research. The application shows it is feasible. The purpose of the paper is to aid special domain information processing and make machine processing more efficient and comprehensive. The research is valuable to ontology development too.

Keywords Web structured information, Information processing, Knowledge base Construction, Semantic resource, Ontology

随着Web的发展,Web中结构化信息与日俱增。如何发现和利用这些信息,日益成为关注的焦点。信息处理需要依赖领域背景知识以及语法语义知识。目前主要通过现有的语义资源和专门领域本体来实现。本文在对Web结构化信息处理对领域知识、语义资源的需求进行分析的基础上,结合结构化数据的特点,综合现有的语义资源构造技术以及本体理论,提出了一个面向特定领域Web信息处理的知识库组成模型,探讨了领域知识库的设计和构建以及构造方法,结合具体的应用检验模型和方法的可行性和有效性。本文的研究主要针对汉语信息。

1 Web结构化信息处理中的汉语资源需求

Web结构化信息处理主要有数据源发现、数据集成、信息检索、数据加工整理等。许多基本操作运算均在不同程度上会依赖背景知识,常见的操作如Web信息自动抽取、包括界面数据模式和元素的自动抽取、检索结果抽取等;Web数据自动分类聚类,如数据源分类、Web查询接口分类、概念词语分类等;多义词义项判断、概念标注;概念或模式的匹配;概念语义范围扩展或收缩;模式和模式成分之间语义关系的确定;概念实例分析等。这些操作需要参照特定的模式知识、语法知识、特定概念语义知识或逻辑推理实现。

另外,Web中存在不同数据源结构异构问题。造成Web语义异构的因素主要有:不同的信息源使用多种术语表示同一概念;同一术语在不同的信息源中表达不同的含义;各信息源使用不同的结构来表示相同或相似的信息;各信息源中的概念之间存在着各种联系,但因为各信息源的分布自治性,这种隐含的联系不能体现出来。语义异构问题的解决需要统一的语义集。

2 汉语语义资源

在中文信息处理领域有影响的汉语语义资源有中科院计算机语言信息工程研究中心开发的知网(HowNet)[1];北大计算语言学研究所的现代汉语语义词典(SKCC)[2]、中文概念词典(CDD)[3]、清华大学计算机科学与技术系的现代汉语语义知识库[4]、台湾中研院的中英双语知识本体词网(SinicaBOW),还有中科院声学所创立的HNC理论中的语义知识库、山西大学的汉语框架语义知识库[5]等。

HowNet是双语常识知识库,通过1500多个义原的组合来描述概念。2005版的HowNet描述了24089个概念,包含中文词语81062个、英文词语76526个。其概念间的关系有上下位、同义、反义、对义、部件-整体、属性-宿主、值-属性、实体-值、相关关系等16种关系。现代汉语语义词典对6.6万余实词进行语义分类,语义搭配限制描述。CCD直接复用WordNet的理论、方法、技术,根据汉语的特点对概念及其关系做了相应调整,目前约包含10万个汉英双语概念。SinicaBOW也是一个双语词汇语义资源,它将WordNet的10万

到稿日期:2008-02-29

郭文宏 博士研究生,主要研究领域为计算机网络信息处理、语义Web、知识工程,E-mail:growh2003@sohu.com;**范学峰** 研究员,博士生导师,主要研究领域为计算机网络与信息处理。

多个概念一一英汉对译，在此基础上将其概念与 IEEE 颁布的 SUMO 本体的节点建立映射关系。现代汉语语义知识库包括现代汉语述语动词机器词典、现代汉语述语形容词机器词典、现代汉语名词槽关系系统、信息处理用现代汉语语义分类词典 4 个资源，对实词进行了详细的句法和语义信息描述，并描述了论旨角色内部的语义组合关系及多项式定语与名词中心词间的语义关系，该词典对 7 万多现代汉语常用动词、形容词、名词的 11 万多个义项进行了分类整理。

语义词典是由一些概念或者概念的同义集合 Lex，并在 Lex 上定义各种关系使之结构化。用二元组〈x，Rx〉来表示概念集 Γ 中的某些概念，其中 $x\in Lex$ 且 Rx 是集合 x 上的关系的集合。不同的语义资源有不同的概念表示形式，概念在《同义词词林》和 HowNet 中表示为义项，在中文概念词典中表示为同义词集。《同义词词林》对汉语词进行了语义分类，整个框架是树形结构。从根部到叶节点语义类逐渐细化，叶节点对应的语义类是同义词群。

利用语义词典概念间显性的关系进行推理外，对树形结构的具有释义文本的语义资源，每个概念和不在同一棵树中的概念也可能有一定的关系，这样就增加了横向联系，从而使整个概念体系呈现为网状结构。根据继承性，下位概念继承上位概念的解释部分，而解释部分本身也存在着一定的层次结构，因此这样就存在着概念的横向关联扩展和纵向关联扩展。横向关联扩展就是扩展到解释部分的上位概念，纵向关联扩展就是扩展到上位概念的解释部分，以此来辅助推理。

词典结构和规模受应用和构造代价的制约，最初设计定位大多是机器翻译应用等信息处理的语法语义分析，也可直接或间接应用于面向 Web 汉语信息处理中。但一般词典是对词汇通用的语义描述，将它直接用于 Web 的语义分析，针对性不强。另外，目前语义资源的术语体系和标注符号体系不统一、规范，不利于资源的共享。资源描述的形式化机读化差，也不利于 Web 应用程序调用。

3 本体知识

本体是共享概念模型的明确的形式化规范说明[10]，表示为五元组 $O=(C,R,Hc,Rel,A0)$，这里 C 是概念集(或类，实体、概念、类型等)，它的元素被称作概念标识符；R 是 $C\times C$ 中关系(或角色)集，其中的元素被称作关系标识符；Hc 是概念集 C 的偏序集，表示概念的层次。Rel 函数表示概念间的非分类关系 $C1\times C2\times\cdots\times Cn-1\rightarrow Cn$；$A0$ 表示本体公理或规则，关系集 R 的偏序集表示关系的层次，公理主要是用来定义词汇的语义和约束用的，对本体中的词汇公理进行定义，主要用一阶谓词逻辑表示，主要描述概念关系的术语公理集和个体实例的断言公理集。本体中的推理包括：计算本体类层次关系，检查概念一致性；计算本体内部关系和隐含关系的合法性；检查实例个体是否是合法的本体实例等。根据概念主题，本体可分为领域本体、通用本体、表示本体、任务本体。

本体是一种能在语义和知识层次上描述系统的概念模型，其目的在于以一种通用的方式来获取领域中的知识，提供对领域中概念的共同一致的理解，从而实现知识在不同的应用程序和组织之间的共享和重用。不同的模型方法、范例、语言和软件工具可借助本体进行转换映射。以本体为基础来指导知识的获取，可提高获取速度和可靠性[8]。本体可解决 Web 结构数据中存在的语义异构问题，但本体建设的现状还处于初级阶段，目前国内权威的针对汉语结构化 Web 信息处理的本体相对较少，构建的原则、方法及其表示等许多方面都没有形成统一的公认的标准。

4 面向 Web 结构化信息处理的知识库模型

在词典等组织方式中，按某种排序方式(如拼音字母序)列出概念及其解释，概念间的联系通过自然语言表达。本体则使用形式化的方法描述概念、实体、过程、属性及其相互关系，并提供了对知识推导的支持。有的词典虽也给出了概念间的联系，但对于联系的类型并没有给出。语义词典和知识本体通常以树或森林组织概念结构，这类语言学资源主要按照概念间结构层次关系组织知识，根据概念间的类属关系、同位关系等来实现语义计算[3,7]。

4.1 知识库组成模型

Web 结构化信息处理所需要的语义资源究竟应是什么样的？研究发现，结构化的数据处理中，经常涉及模式间的关系、模式元素与模式整体的关系、模式内部各成分间的语义约束关系，因此数据关系模型知识及特定领域中模式概念结构知识对此类问题的解决很有用。其次，结构化 Web 信息大多是语言文本，经常需要对其进行词法语义分析，相关语法知识对此类问题的处理也很必要。另外，Web 数据源存在结构异构问题，缺乏统一的语义集，目前利用本体知识来解决。通过以上分析，本文基于现有的本体理论、语言学词汇语义网构建理论以及关系数据库理论，提出了一个面向特定领域结构化信息处理的汉语领域知识库组成模型。该模型除了描述领域概念的属性、关系外，同时侧重于描述结构模式实体及其相关知识。另外，结合本体概念描述和语言学词汇词条定义的特点，加入更多词法语法知识。为计算机对语言的自动分析提供支持。该模型融合了领域世界的概念及语义关系、数据模式及其相关关系、汉语资源的词法语法知识，在具体实现中，这 3 部分根据需要可进行适当内聚融合，也可保持松耦合的关系。极端情况是 3 个部分融合为一体，整个领域知识库可看作对领域本体的扩展；另一个极端是保持 3 个部分为 3 个相对独立的知识库。知识库组织构造演化示意图如图 1。

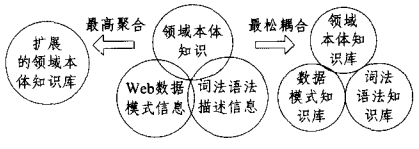

图 1 Web 结构化信息处理的汉语知识库构造组织演化示意图

该知识库模型形式化表示为 $KB=(C,S,G,R)$，其中 C 表示领域概念集，S 表示数据模式集，G 表示概念、模式的词法语法信息，R 表示包括与模式相关关系的语义关系集。该模型在注重描述世界本原的同时，增加了模式功能表现形式，模式本身也被看作一种客观存在。该模型对本体的概念粒度和语义关系进行扩展，引入数据模式和与其相关的关系(模式元素和模式整体关系，模式间的约束关系等)，作为特殊的语义关系。另外，结合本体概念描述和语言学词汇词条定义的特点，增加了概念和术语表达语法特性，以详尽描述概念在模式语境中的表现和内涵个性特征。该知识库模型提供以下内容：(1) 领域词汇表及词法语法知识；(2)领域知识的显式表

示；(3) 目标领域世界的元模型，明确地表示人们对目标世界的共同理解。

4.2 知识库构建方法

借鉴骨架法和应用驱动的本体构建方法[6]，Web 领域知识库构建示意图如图 2 所示。

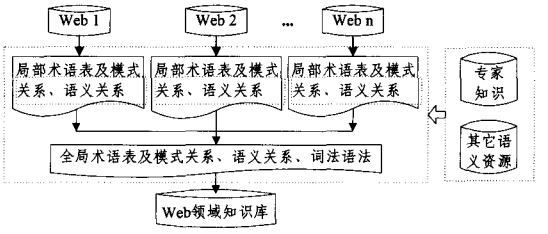

图 2 面向 Web 信息处理的领域知识库构造示意图

首先收集特定领域中各 Web 数据源信息，分析资源数据、数据存储方式以及数据接口模式语义信息，建立各 Web 数据源内部模式和概念术语列表，定义这些概念的性质和属性，确定模式的层次结构，明确概念间的关系；其次利用专家知识和其它语义资源，如现有本体、语义词典、已有的模式知识库等，分析各局部概念术语之间的关系，抽取全局术语列表、全局模式和语义关系集；再次根据全局概念和关系进行集成，使之成为一个体系；最后定义并保存全局本体与各局部本体之间的映射关系[8,9]。

4.3 知识库实现

基于以上理论针对 Deep Web 研究构建了多个领域的知识库，领域知识库由不同的概念、实体及其间的关系、与之对应的词条组成，且显式地增加了模式类概念和元素类概念。元素类概念和普通概念的唯一区别就是其可作为特定模式的元素。模式类的描述主要有模式名称、关键元素集、其它元素集、子模式集、父模式集、其它关联模式集、元素概念在模式空间的相对位置信息、模式实例等。元素类的描述主要有元素名称、元素类型、元素解释说明信息如义项、元素与其相关模式间关系、元素概念间关系、元素实例等。概念间的关系通过属性和显式定义的语义关系来描述。典型的关系有同义或等价关系、类属或上下位关系、合成或部分整体关系等。模式间的关系 $R_{S\times S}$＝{等价，互斥，相容，父模式，子模式，关联，…}；元素间关系 $R_{E\times E}$＝{互斥，相容，同义，相关，空间位置先后顺序关系，时间先后顺序关系，…}；概念间关系 $R_{C\times C}$＝{同义，超类，子类，成员，部分，…}，这里 $R_{E\times E}\supset R_{C\times C}$；实例间的关系 $R_{I\times I}$＝{相关，合成，类属，…}；属性间的关系 $R_{A\times A}$＝{类属关系，…}。另外还有概念与个体关系 $R_{C\times I}$＝{实例关系（*is-A*）}，概念与其属性间的特征关系等。属性有不同的侧面，诸如属性值的类型、属性值的取值范围、属性值的数量及子元素个数等其它特征。模式与其元素间的关系 $R_{S\times E}$ 通过关键元素、其它元素、上/下位模式集等属性来表示。元素语法信息有概念释义信息、概念表示形式、词性、配价数、数量单位、语义搭配限制、拼音、对应英语单词等。

图 3 是针对航空查询的知识库中单程航班查询模式知识关系示意图，图中的单程航班查询的关联模式集中，单程航班查询与往返航班查询、联程航班查询是互斥的模式关系，与国内航班查询、国际航班查询是相容的模式关系。

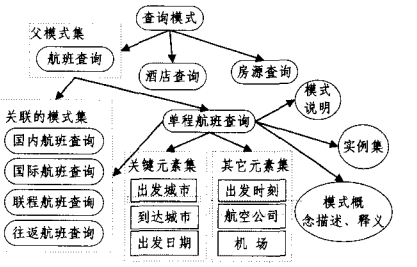

图 3 单程航班查询概念模式关系示意图

图 4 是概念起飞城市的关系示意图。航空查询知识库是建立的知识库之一，另外还建立了酒店查询、房源查询等领域知识库。

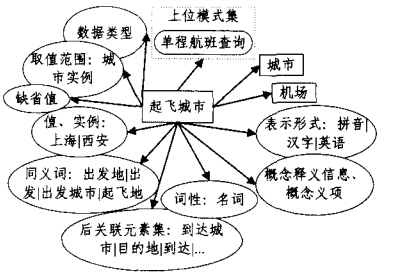

图 4 起飞城市在知识库中关系示意图

利用该知识库模型，推理机制可根据不同关系的相关性质确定泛化的推理规则，如类间的传递性规则、数据属性继承规则、实例传递归属规则、父子类逆关系规则；实例间的对象属性继承规则、数据属性继承规则；属性间的传递性规则、属性外延规则、属性外延具体化规则、父子属性逆关系规则等。然后考虑对象属性的定义域与值域，抽取出满足具体应用的规则。这样扩展了对象属性和断言事实，进而拓展了知识库知识。

结束语 本文在对现有信息处理用汉语资源及技术整合的基础上，提出了针对结构化 Web 信息的知识库构建模型并加以实现验证。研究认为，显式地在知识库中增加特定领域的数据模式知识，能够满足结构化 Web 信息处理中多方面的语义计算和操作的需要，在实际应用中很有价值。此外，在知识库中增加词语或结构的语法信息，有助于提高计算机理解处理信息的准确性和效率。基于该知识库模型，语义计算时可以充分利用语言学特性、模式语义约束等特性。该模型旨在辅助实现特定领域的结构化信息处理，使计算机对语言文本信息处理更全面、更有效，对本体的深入研究也有一定的参考价值。工程性是知识库构造的天然属性，基于本文提出的模型中知识的表示、存储、组织有待进一步深入研究和探索。

参 考 文 献

[1] 董振东，董强. 知网(Hownet)[R]. http://www.Keenage.com

[2] 王惠，詹卫东，俞士汶. 现代汉语语义词典的结构及应用. 语言文字应用，2006(1)：136-141

[3] 于江生，俞士汶. 中文概念词典的结构. 中文信息学报，2002，16(4)：13-21

[4] 陈群秀. 一个现代汉语语义知识库的研究和实现//曹右琦，孙茂松. 中文信息处理前沿进展. 清华大学出版社，2006：172-182

[5] 刘开瑛，由丽萍. 汉语框架语义知识库构建工程//曹右琦，孙茂松. 中文信息处理前沿进展. 清华大学出版社，2006：64-72

[6] William S, Austin T. Ontologies[J]. IEEE Intelligent Systems, 1999, 1(2): 18-19

[7] Gruber C T R. A translation approach to portable ontologies [J]. Knowledge Acquisition, 1993, 5(2): 199- 220

[8] Ushold M, Gruninger M. Ontologies: Principles, Methods and Applications[J]. Knowledge Engineering Review, 1996, 11(2): 93-126

[9] Rodriguez M A, Egenhofer M J. Determining Semantic Similarity Among Entity Classes from Different Ontologies [J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(2): 442-456

[10] Studer R, Benjamins V R, Fensel D. Knowledge Engineering: Principles and Methods [J]. IEEE Transactions on Data and Knowledge Engineering, 1998, 25: 161-197

[11] Chandrasekaran B, et al. What are Ontologies, and Why do We Need Them[J]. IEEE Intelligent Systems, 1999, 14(1): 20-26

(上接第 147 页)

(3)找出适应度最差的粒子，根据式(6)、(7)进行下一次进化，其余的根据式(4)、(5)进行下一次进化；

(4)更新各个粒子的个体历史最好适应值和个体历史最好位置；更新全群历史最好适应值和全群历史最好位置；

(5)若满足停止条件(迭代次数超过最大允许迭代次数)，搜索停止，输出全群历史最好适应度值；否则，返回步骤(2)继续搜索。

4 算例

为了说明本文所提出的模型和算法的可行性和有效性，这里使用 Markowitz[1] 给出的一个实例。另外，粒子群算法中相关参数设置为：种群的规模为 30，迭代次数为 3000。

现假设模型(3)中 $t_g=0.3$，$t_c=0.035$，$t_s=0.001$，$\boldsymbol{x}^0=(0,0,0,0,0,0)$，$\boldsymbol{u}=(1,1,1,1,1,1,1,1,1,1)$，$\boldsymbol{l}=(0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01)$。最后计算结果见表 1。

表 1 最优投资组合

r_0	0.08	0.085	0.09	0.095	0.10
x_1	.139689	.1841450	.0617210	.1926300	.2297760
x_2	.0168303	.2108090	.0583260	.0810378	.0217496
x_3	09804450	.1814870	.0681990	.0993568	.0649846
x_4	.0271335	.0263139	.1128560	.2093570	.0494227
x_5	.0979808	.1971000	.1478830	.0136977	.113534
x_6	.0118057	.0235325	.1050080	.0836787	.0195537
x_7	.2873780	.0363526	.1069840	.0532802	.0625928
x_8	.0737792	.0519482	.2204450	.1844840	.3268090
x_9	.2473540	.0883083	.1185730	.0824731	.1115740
risk	.0102865	.0114623	.0122722	.0123707	.0137749

由表 1 中可以看出，随着期望收益率的增加，证券投资组合的风险也在不断地增大，这符合实际情况。另外，如果投资者对已得到的投资组合策略不满意，可以通过调整 r_0 获得更多的投资策略。

最后，为了说明 IPSO 算法优于 PSO 算法的有效性，本文从以下 4 个方面进行比较：成功率(精度为 10^{-5})，最好目标函数值，最差目标函数值，平均目标函数值。这里，我们以 $r_0=0.08$进行试验，算法运行 10 次，实验结果见表 2。

表 2 算法比较结果

	IPSO	PSO
成功率(%)	100	10
最好值	0.0102865	0.0242099
最差值	0.0156033	0.0242099
平均值	0.013083747	0.0242099

从表 2 中可以看出：对于我们所提出的问题，IPSO 算法比标准 PSO 算法具有更强的寻优能力，可以获得更精确的最优解。例如，IPSO 算法所得到的目标函数的平均值为 0.013083747，而 PSO 所得到的目标函数的平均值却只有 0.0242099。

结束语 本文研究了更加符合我国现实投资环境的投资组合选择问题，提出了具有交易成本和投资数量限制的投资组合模型，并进一步设计了一种改进的粒子群算法求解该模型。实证结果表明：我们所提出的模型和方法是有效的。

参 考 文 献

[1] Markowitz H M. Portfolio Selection. Journal of Finance, 1952, 7: 77-91

[2] Chang T-J, Meade N, Beasley J, et al. Heuristics for Cardinality Constrained Portfolio O-ptimization. Computers and Operations Research, 2000, 27: 1271-1302

[3] Yu L, Wang S Y, Lai K K. Neural network-based mean-variance-skewness model for portfolio selection. Computers & Operations Research, 2008, 35(1): 34-46

[4] Crama Y, Schyns M. Simulated Annealing for Complex Portfolio Selection Problems. Euopean Journal of Oprational Research, 2003, 150: 546-571

[5] 林丹，李小明，王萍. 用遗传算法求解改进的投资组合模型[J]. 系统工程，2005，23(8)：68-72

[6] 荣喜民，李楠. 考虑完整交易费用的组合证券投资求解[J]. 数学的实践与认识，2007，37(10)：22-27

[7] 周洪涛，刘康泽. 摩擦市场条件下的双目标投资组合模型模糊优化[J]. 数学的实践与认识，2007，37(7)：27-32

[8] Kennedy J, Eberhart R C. Particle Swarm Optimization // Proceedings of IEEE International Conference on Neural Networks. Piscataway, 1995, 10: 1942-1948

[9] Eberhart R, Kennedy J. A New Optimizer Using Particle Swarm Theory // Proc. 6th Int Sympossum on Micro Machine and Human Science. 1995, 11: 39-43

[10] Arnott R D, Wanger W H. The Measurement and Control of Trading Costs. Financial Analysts Journal, 1990, 46 (6): 73-80

[11] Yoshimoto A. The Mean-Variance Approach to Portfolio Optimization Subject to Transaction Costs. Journal Research Society of Japan, 1996, 39: 99-117

[12] Liu M M, Gao Y. An algorithm for portfolio selection in a frictional market. Applied Mathematics and Computation, 2006, 182 (2): 1629-1638

[13] Yang J M, Chen Y P, Horng J T, et al. Applying Family Competition to Evolution Strategies for Constrained Optimization. Lecture Notes in Computer Science, Springer-Verlag, 1997, 1213: 201-211