基于Ontology的信息抽取研究综述

陈 静,朱巧明,贡正仙
(苏州大学 计算机科学与技术学院,江苏 苏州 215006)

摘 要:信息抽取研究旨在为人们提供更有力的信息获取工具,以应对信息爆炸带来的严重挑战。Ontology作为领域知识的共同理解,能有效地解决现在信息抽取所面临的主要挑战——知识工程的瓶颈问题。文中详细介绍了本体的定义和建模语言,分析了现有基于本体信息抽取的几种典型方法,得出了其所存在的主要问题。

关键词:信息抽取;本体;知识获取
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2007)10-0084-03

Overview of Ontology-Based Information Extraction

CHEN Jing, ZHU Qiao-ming, GONG Zheng-xian
(Department of Computer Science and Technology, Soochow University, Suzhou 215006, China)

Abstract: The research on information extraction aims at providing more powerful information access tools to help people overcome the problem of information overloading. As the common knowledge of a special domain, ontology could deal with the challenging task in information extraction: the bottleneck of knowledge engineering. In this paper, presents the definition and modeling language in detail. And discusses the popular methods used in information extraction based on ontology. At last put forward some problems in those systems.

Key words: information extraction; ontology; knowledge acquisition

0 引 言

为了应对信息爆炸带来的挑战,迫切需要一些自动化的技术帮助人们在海量信息中迅速找到自己真正需要的信息,信息抽取技术应运而生,成为目前自然语言处理领域的研究热点。

信息抽取是以一个未知的自然语言文档作为输入,产生固定格式、无歧义的输出数据的过程。这些数据可以直接向用户显示,也可作为原文信息检索的索引,或存储到数据库、电子表格中,以便于以后的进一步分析。

传统的信息抽取系统多采用基于模板和模式匹配,或者是采用基于统计的学习方法[1]。这些方法都需要在前期进行大量的手工标注训练文本,然后对训练文本进行学习。但是训练文本不可能覆盖整个领域内出现的所有语言习惯。另外,传统的信息抽取虽然能抽取出实体,但是缺乏领域知识来识别抽取实体之间的关系[2]。因此在信息抽取任务中引入相应的领域知识——领域本体来指导抽取过程,将能有效地提高信息抽取的性能。

收稿日期:2006-12-14
基金项目:江苏省高技术研究项目(BG2005020);江苏省教育厅自然基金(04KKB320134)
作者简介:陈 静(1982-),女,湖北汉川人,硕士研究生,研究方向为中文信息处理;朱巧明,教授,研究方向为中文信息处理、网格计算、分布式计算。

1 本体定义

Ontology最早是一个哲学的范畴,后来随着人工智能的发展,被人工智能界赋予了新的定义。最初人们对Ontology的理解并不完善,这些定义也处在不断的发展变化中,比较有代表性的定义如表1所示。

表1 本体概念的发展过程[3]

范畴	提出时间/提出人	定义
哲学	公元前/亚里士多德	客观存在的一个系统的解释和说明,客观现实的一个抽象本质
计算机	1991/Neches[4]等	给出构成相关领域词汇的基本术语和关系,以及利用这些术语和关系构成的规定这些词汇外延的规则的定义
	1993/Gruber[5]	概念模型的明确的规范说明
	1997/Borst[6]	共享概念模型的形式化规范说明
	1998/Studer[7]	共享概念模型的明确的形式化规范说明

尽管大家对本体没有一个明确统一的定义,但是从内涵上来看,不同研究者对于本体的认识是统一的,都把本体当作是领域(可以是特定领域的,也可以是更广的范围)内部不同主体(人、机器、软件系统等)之间

进行交流(对话、互操作、共享等)的一种语义基础,即由本体提供一种明确定义的共识。

Ontology 是描述概念及概念间关系的概念模型,通过概念之间的关系来描述概念的语义。作为一种有效表现概念层次结构和语义的模型,Ontology 被广泛地应用到计算机科学的众多领域:文本分类、信息检索、异构信息系统集成和语义 Web 等。

2　本体建模语言

Ontology 的目标是捕获相关的领域的知识,提供对该领域知识的共同理解,确定该领域内共同认可的词汇,并从不同层次的形式化模式上给出这些词汇(术语)和词汇之间相互关系的明确定义。Ontology 是一种组织知识的艺术,为研究如何用 Ontology 来组织知识,文献[5]用分类法,归纳出组成本体的 5 个基本的建模元语为:(C, I, R, F, A)。其中:

C:类(classes) 或概念(concepts),指任何事务,如工作描述、功能、行为、策略和推理过程。从语义上讲,它表示的是对象的集合,其定义一般采用框架(frame)结构,包括概念的名称,与其他概念之间的关系的集合,以及用自然语言对概念的描述。

I:实例(instances),代表元素。从语义上讲实例表示的就是对象。

R:关系(relations),定义在概念集合上的关系集合。在领域中概念之间的交互作用,形式上定义为 n 维笛卡儿积的子集:$R: C_1 \times C_2 \times \cdots \times C_n$。如子类关系(subclass - of)。在语义上关系对应于对象元组的集合。

F:函数(functions),定义在概念集合上的函数集合,对于任意一个 $f \in F$,f 是一个 $n-1$ 元有序序列到一个概念的映射 $f = (C_1, C_2, \cdots, C_{n-1}) \rightarrow C_n$。如 Mother - of 就是一个函数,mother - of(x, y) 表示 y 是 x 的母亲。

A:公理(axioms),代表永真断言,如概念乙属于概念甲的范围。

当然本体的建模语言并不是说必须参照这样完备的集合,对于一些轻量级的本体仅由 C, R, I 组成。

3　基于本体的信息抽取的研究现状

领域本体由领域内的相关概念、属性、关系、约束及术语或实例等构成。基于本体的信息抽取主要是利用领域本体对领域内数据的描述信息来实现抽取,因此领域本体构建的好坏将直接影响到信息抽取的性能。

目前,基于 ontology 的信息抽取的方法已经有了不少的研究,通常来说基于 Ontology 的信息抽取方法主要分为以下两类。

3.1　知识工程的方法

由专家对 Ontology 进行分析、调整而人工制定规则、模板。文献[8]中根据选定的训练集中的数据来确定本体中出现的概念和关系,建立本体;手工统计概念和关系中出现的关键字,书写正则表达式来表示抽取规则,然后根据规则进行抽取。因为本体的构建和规则的制定是建立在特定的训练集上的,实际的抽取过程中如果抽取的文本结构和表述方式发生变化时,对于抽取的结果影响非常大。

文献[9]把语法分析和 Ontology 结合起来,利用领域 Ontology 里的概念、关系、关键字自动生产抽取规则,然后对文章、句子的语法结构进行分析,然后利用规则对文档进行标注与抽取。生产的规则中的关键字也是由手工统计,覆盖范围有限。

文献[2]中介绍了 Artequakt 系统框架中的信息抽取模块,它使用现有的 IE 工具和技术来实现实体标注,然后从本体中推导和决定实体间可能存在的二元关系,进行实体关系的识别,进而进行知识抽取。另外,还使用了基于字典的字典术语机制来改善抽取结果,即利用 WordNet 的字典链(同义关系、上位关系和下位关系)来降低语言使用习惯的变化对本体关系识别和抽取数据的影响。

类似的研究可见文献[10~12]。

3.2　自动训练方法

给出根据本体中的概念进行标注的例子文档集,通过机器学习的方法来推导模板和模板的自动填充知识库和规则。

文献[13]介绍了一个用于知识抽取的语义标注工具,它主要包括三个部分。本体标注组件提供了一个用户接口,方便用户根据本体的概念和关系来对训练集中的信息片断进行标注;学习组件使用现有的学习工具(Crystal from the University of Massachusetts)对标注好的训练语料进行学习得到规则;最后,信息抽取模块根据学习得到的规则进行抽取。另外,利用本体中的实例来消除抽取过程中的信息二义性问题。例如,"X 被 Y 授予一定数量的金钱奖励",在抽取时,X 既可以作为一个组织名也可以作为一个项目组的名称,通过将 X 与本体中实例进行匹配,可以得出 X 所属的本体概念。

4　基于本体的信息抽取的不足

虽然目前已经有很多利用本体进行信息抽取的系统,也取得了一些效果,但是这些系统很少能发挥领域本体的在信息抽取过程中的全部潜能。通过对现有系

统的研究,发现基于本体的信息抽取研究存在以下几个不足之处。

4.1 本体建立不够完善

因为基于本体的信息抽取是利用领域本体对领域内数据的描述信息进行抽取的,所以本体的好坏与否将直接影响抽取的性能。文献[8]和文献[9]都对本体进行了简化的处理,使用OSM模型来描述本体,其建立的本体仅仅包含本体建模语言5个组成元语中的C与R,文献[2]与文献[13]中的本体也仅包含C、R和I。

虽然本体在信息抽取过程中有着举足轻重的地位,但是因为本体的建立需要领域专家的参与且需要耗费大量的人力和物力,特别是领域内实例术语的获取如果依靠手工获得,是不现实的,因此现在大量的研究来自动构建本体,自动获取本体的概念和关系,并利用信息抽取技术来自动获取本体中的实例[12~15]。

4.2 没有充分发挥本体在信息抽取过程中的潜力

用来抽取的本体多数仅仅由C与R建模元素构成,那么在抽取的过程中也都只有利用到本体的概念和概念间的关系,本体在抽取的过程中并没有发挥本体的全部潜能。文献[2,8~12]在抽取中都仅仅使用了本体的概念和概念间的关系(C,R);文献[13]除了使用了本体的概念和概念间的关系之外,还利用了本体中的实例(I)来消除抽取过程中的歧异,但是实例的利用仅仅在实体识别之后,用来判断实体所属的本体概念。本体中的实例表示了领域内的一些特殊的概念术语,如果能把实例用于实体的发现也将大大提高抽取的性能。其次,本体中还包含概念属性、概念的约束等信息也能对抽取过程进行指导。

另外,通过前面的分析发现,无论是领域本体还是在信息抽取的过程中,本体建模语言中的F和A都没有涉及到,而F和A却是本体具有推理功能的重要前提。本体的推理能够对抽取到的消息进行推理组合,把隐含在事件中的信息通过推理过程进行明确。例如:本体中存在两个函数:Father(x,y)表示x是y的父亲、Grandfather(m,n)表示m是n的祖父;存在公理:if Father$(x,y)$and Father(y,z)then Grandfather(x,z)。如果抽取得到下面的事实:张三是张四的父亲,张四是张五的父亲,那么根据上面的公理可以推理出张三是张五的祖父。

4.3 移植性不高

现有的基于本体的信息抽取系统都集成了一些现有的传统的信息抽取方法,在前期都需要进行大量的手工标注或规则的获取过程,因此也存在传统的信息抽取方法中的移植性问题。另外,目前的领域本体构建的自动化程度不高,本体构建费时费力也给系统的移植性带来了一定的困难。

5 总 结

目前,基于本体的信息抽取技术的研究虽然很多,但是仍然属于一个探索的阶段,并没发挥本体的全部潜能,其根本的瓶颈是本体的构建问题。目前本体的自动构建技术还很不成熟,领域知识的自动获取还依赖于信息抽取技术,因此本体的自动构建技术将与信息抽取技术相互促进,共同前进,基于本体的信息抽取技术将有很大的发展空间,成为研究的热点。

参考文献:

[1] 李向阳,苗 壮.自由文本信息抽取技术[J].情报科学,2004,22(7):815-821.

[2] Alani H,Kim S,Millard D E,et al. Automatic Ontology Based Knowledge Extraction from Web Documents[J]. IEEE Intelligent Systems,2003,18(1):14-21.

[3] 邓志鸿,唐世渭.Ontology研究综述[J].北京大学学报:自然科学版,2002,38(5):730-738.

[4] Neches R,Fikes R E,Gruber T R,et al. Enabling Technology for Knowledge Sharing[J]. AI Magazine,1991,12(3):36-56.

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

[6] Borst W N. Construction of Engineering Ontologies for Knowledge Sharing and Reuse[D]. Enschede: University of Twente,1997.

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

[8] Embley D W. Ontology-based Extraction and Structuring of Information from Data-Rich Unstructured Documents[C]// Proceedings of Conference on Information and Knowledge Management. Bethesda,Maryland,USA:[s. n.],1998:52-59.

[9] 陈 兰,左志宏,熊 毅,等.一种新的基于Ontology的信息抽取方法[J].计算机应用研究,2004,21(8):155-157.

[10] Snoussi H,Magnin L,Nie J Y. Heterogeneous Web Data Extraction using Ontology[C]//In Third International Bi-Conference Workshop on Agent-oriented information systems (AOIS-2001). Montréal,Canada:[s. n.],2001.

[11] Reyle U,Saric J. Ontology Driven Information Extraction [C]//In Proceedings of the nineteenth Twente Workshop on Language Technology. Enschede: University of Twente,2001.

[12] Honavar V,Silvescu A,Reinoso-Castillo J. Ontology Driven Information Extraction and Knowledge Acquisition from Het-

(下转第91页)

微大一些的块。动态调度的灵活性最大,但是在调度出错时,造成的性能损失也最大。对于平衡良好的代码,应避免使用非常小的块。本例子中调度块大小为128明显是很好的选择。而对于那些不平衡的代码,正确的块大小应该在运行较少的块和达到更佳的线程平衡之间做一个权衡。引导调度使用较小的块作为限制时运行最佳、灵活性最大。限定为较大的块大小时,运行的时间将很长。引导调度的上限是循环总大小的二分之一,对于平衡良好的代码,应该相当于在一次运行中将循环拆分为相等几部分的静态情况。

3 结　语

OpenMP 是针对共享地址空间并行计算机提供的并行计算库,目前 Microsoft 和 Borland 公司都有相应产品支持 OpenMP2.5。使用 OpenMP 不必写诸如 CreateThread 之类的线程管理代码,编写多线程程序简便高效,而且 OpenMP 提供了丰富的指令,对于同步共享变量、合理分配负载等任务,都提供了有效的支持。不过 OpenMP 也存在着一些不可避免的缺点:第一,OpenMP 主要以预编译指令(#pragma)实现多线程并行,所以在单核机器上编译的程序在多核机器上运行时无法体现多核的优势;第二,OpenMP 对编译器要求比较高,一般要求 Microsoft Visual Studio 2005 或者需要 Intel 编译器。不过长远来讲,OpenMP 的优势是明显的。Intel 技术官曾说"今后的处理器发展是内部优化与集成多核而不是单纯地提升处理器的频率,采用多线程的软件也将会是今后软件的主流。"充分发挥多核处理器的优势使软件性能最优化,给中国带来一个极大的机遇。开发 OpenMP 软构件库需要大量的人力资源,这项工作适合在中国进行。OpenMP 库在未来的经济效益可以同微软的操作系统相比,后者为用户顺利使用 PC 提供工具软件支持,前者为顺利使用多处理器提供库函数支持。世界正在进入多处理器时代,OpenMP 库将成为程序员必不可少的工具。

参考文献:

[1] 陈国良.并行算法实践[M].北京:高等教育出版社,2004.

[2] Quinn M J. Parallel programming in C with MPI and OpenMP [M]. 北京.清华大学出版社,2005.

[3] 赖建新,胡长军,赵宇迪,等. OpenMP 任务调度开销及负载均衡分析[J]. 计算机工程,2006,18:(sup)58-60.

[4] Grama, Ananth. Introduction to parallel computing[M]. 北京:机械工业出版社,2003.

[5] Foster I, Designing and building parallel programs[M]. 北京:机械工业出版社,2002.

[6] Andrew. Multithreading parallel and distributed programming [M]. 北京:高等教育出版社,2002.

[7] Malyshkin V. Parallel computing technologies[C]//8th international conference, PaCT 2005. Krasnoyarsk, Russia, 2005. Berlin; New York: Springer, 2005.

[8] 杨淑莹. VC++图像处理程序设计[M]. 北京:清华大学出版社,2003.

[9] Dongarra J. Parallel computing programming[M]. 北京:电子工业出版社,2005.

[10] Wilkinson B, Allen M. Techniques and applications using networked workstations and parallel computers[M]. 北京:机械工业出版社,2005.

(上接第 83 页)

川大学学报:工程科学版,2005,37(3):118-122.

[14] Govindan R, Tangmunarunkit H. Heuristics for internet map discovery[C]//Proceedings IEEE INFOCOM. [s. l.]: [s. n.], 2000:1371-1380.

[15] 姜　誉,胡铭曾,方滨兴,等. 一个 Internet 路由器级拓扑自动发现系统[J]. 通信学报,2002,23(2):54-62.

[16] 杨家海,任宪坤,王沛瑜. 网络管理原理与实现[M]. 北京:清华大学出版社,2000.

[17] Peterson L L, Davie B S. Computer networks: A Systems Approach[M]. 3rd Edition. USA: Elsevier Science, 2003.

[18] RekhterY. An architecture for IP address allocation with CIDR [S]. RFC1518, 1993.

(上接第 86 页)

erogeneous, Distributed Biological Data Sources[C/OL]// Proceedings of the IJCAI2001 Workshop on Knowledge Discovery from Heterogeneous, Distributed, Autonomous, Dynamic Data and Knowledge Sources. 2001. http://www.cs.iastate.edu/honavar/Papers/ijcaiworkshoppaper.pdf.

[13] Vargas-Vera M. Knowledge Extraction Using an Ontology-Based Annotation Tool [C]//Workshop on Knowledge Markup & Semantic Annotation. [s. l.]: ACM Press, 2001:5-12.

[14] Maedche A, Neumann G, Staab S. Bootstrapping an Ontology-based Information Extraction System[C]//Intelligent Exploration of the Web, Studies in Fuzziness and Soft Computing. Heidelberg: Physica-Verlag GmbH, 2003:345-359.

[15] Yeh C L, Su Y C. Web Information Extraction for the Creation of Metadata in Semantic Web[C]//Proceedings of ROCLING. Tainan, Taiwan: [s. n.], 2005.