Ontology 自动创建中实例学习的研究*

刘贺欢，刘椿年

(北京工业大学 计算机学院 多媒体与智能软件技术北京市重点实验室，北京 100022)

摘 要：Ontology 近年来受到信息科学领域的广泛关注，其重要性已在许多方面表现出来并得到广泛认同。自动创建领域 Ontology 可以克服手工方法的不足，成为当前的研究热点之一；实例(Instance)是 Ontology 的重要组成元素，从领域文档集中学习实例是自动创建领域 Ontology 的关键之一。研制的一个领域 Ontology 的自动生成系统(OntoAGS)能够通过领域文档集自动地创建该领域的 Ontology，OntoAGS 系统的实例学习是基于模式匹配的算法。实验表明，与当前较流行的 Ontology 半自动生成系统 text-to-onto 相比，OntoAGS 通常能学习出更多实例，且准确率较高。

关键词：自动创建领域 Ontology；模式匹配；实例学习

中图法分类号：TP39 **文献标识码**：A **文章编号**：1001-3695(2006)03-0038-03

Instance Learning in Automatic Construction of Domain Ontology

LIU He-huan, LIU Chun-nian

(Beijing Municipal Key Laboratory of Multimedia & Intelligent Software Technology, College of Computer Science, Beijing University of Technology, Beijing 100022, China)

Abstract: The importance of Ontology is being recognized in various research and application areas. Constructing domain Ontology takes too much time and manpower. Hence, it is of great significance to do some research on constructing domain Ontology automatically. Instance is an important element of Ontology. Learning instance from domain corpus is a key issue in automatic construction of domain Ontology. A domain Ontology construction system OntoAGS has been developed. Given some domain corpus, OntoAGS can construct domain Ontology automatically. The instance learning method of this system is based on pattern matching algorithm. Experiments show that OntoAGS can generally learn more instances and more accurately than text-to-onto system which is a popular Ontology semi-automatic construction system.

Key words: Automatic Construction of Domain Ontology; Pattern Matching; Instance Learning

Ontology 最早是一个哲学上的概念，20 世纪 80 年代，Ontology 应用到计算机领域。1998 年 Studer 等人对 Ontology 概念进行了深入研究，认为 Ontology 是“共享概念模型的明确的形式化规范说明[1]”。Ontology 可以手工或半自动创建，但创建的时间长、工作量大。自动创建 Ontology 的领域随之发展起来。该领域致力于克服手工和半自动创建 Ontology 的不足，从而受到国内外众多研究人员的关注。实例(Instance)是 Ontology 的重要组成元素，实例学习是自动创建 Ontology 的关键技术之一。实例学习尚处于研究探索阶段，目前还没有比较成熟的算法。现在已经开发出了一些 Ontology 创建工具，如 OntoLingua, OntoSaurus, WebOnto, Protégé[2], text-to-onto[3] 等，这些工具都是手工或半自动创建 Ontology。我们开发了一个 Ontology 自动生成系统 OntoAGS，该系统可以从给定领域的文档集中自动创建该领域的 Ontology，OntoAGS 系统的实例学习是基于模式匹配的算法。实验证明，OntoAGS 系统学习的实例在准确率和数量方面一般都优于 text-to-onto 系统。

收稿日期：2005-01-21；修返日期：2005-06-20

基金项目：国家自然科学基金资助项目(60173014)；北京市自然科学基金资助项目(4022003)

1 Ontology 学习简述

Ontology 的基本元素包括概念(Concept)、实例(Instance)和关系(Relation)等。概念的含义很广泛，可以指任何事物，如工作描述、功能、行为、策略和推理过程等。从语义上分析，概念表示的则是对象的集合[4]；实例是组成概念的成员[5]。这里的关系是指概念之间、实例之间，以及概念与实例之间的关系。自动创建 Ontology 涉及到对组成 Ontology 元素的学习，因此，Ontology 学习主要包括概念学习、实例学习和关系学习等。实例学习就是从给定的领域文档集中自动地学习出该领域的实例。下面主要介绍 OntoAGS 系统的实例学习算法，该系统的实例学习是基于模式匹配的算法。

2 OntoAGS 系统的实例学习

我们开发了一个 Ontology 自动生成系统 OntoAGS，该系统能够从给定的领域文档集中自动地创建该领域的 Ontology。通过分析大量的句子，我们可以总结出包含实例的各种句子结构，学习实例的模式是根据这些句子结构定义的，然后把这些模式运用到实例学习的过程中，从而从领域文档集中自动地学习出所研究领域的实例。

2.1　基于模式匹配的实例学习算法

模式匹配是指在目标串中查找与模式串相同的子串的过程。模式匹配的含义可以直观地从下例中看出来。

例 1: 设目标串为 abcdadcab; 模式串为 ab, 则模式匹配后查找到目标串中与模式串相同的子串的首位置是 1 和 8。

OntoAGS 系统的实例学习采用基于模式匹配的算法, 该算法的主要步骤如算法 1 所示。

算法 1:

(1) 构造领域文档集。

(2) 利用 QTag[6] 对领域文档进行词性标注。

(3) 定义学习实例的模式。

(4) 把词性标注后的领域文档中的每句话与所定义的模式进行匹配。如果匹配成功, 则把实例从此句话中取出, 并还原为词干。进行下一句话的匹配。

2.2　构造领域文档集和词性标注文档

算法 1 的第一步是建立学习实例的领域文档集, 领域文档集中包含所要进行实例学习的领域文档。第二步是利用 QTag 标注对领域文档集中的领域文档进行词性标注, 并把标注后的结果作为模式匹配算法的目标串。不同的词性标注开发者都会定义自己的词性标注的标签, 表 1 给出了 QTag 的部分词性标注标签。

表 1　QTag 的部分标签与含义

标　签	含　义
BEDR	were
BEDZ	was
BER	are
BEZ	is
CC	连接词
DT	限定词(a, the, this, that)
HV	have
HVZ	has
JJ	形容词
NN	普通名词的单数形式
NNS	普通名词的复数形式
NP	专有名词的单数形式
NPS	专有名词的复数形式

下面的例子说明了运用 QTag 对句子进行词性标注的过程和结果。

例 2: 给定的句子为 MIT is a famous university, 运用 QTag 对该句话进行词性标注的过程和标注结果如下:

(1) 标注过程

MIT 被标注为 NP, is 被标注为 BEZ, a 被标注为 DT, famous 被标注为 JJ, university 被标注为 NN。

(2) 标注结果

标注结果被存储为如下格式:

NP MIT BEZ is DT a JJ famous NN university…

其中, 每一个单词的词性标签放在该单词的前面, 并用空格隔开标签及单词, 该单词与下一个单词的词性标签之间用制表符分隔。例如 RB 与 there 之间用空格分隔, there 与 BER 之间用制表符分隔。

2.3　模式定义

通过分析大量的句子, 我们总结出包含实例的各种句子结构, 学习实例的模式是根据这些句子结构定义的。随着人工分析出包含实例的句子结构数量的增加, 实例模式随之将不断地进行丰富。OntoAGS 系统定义了许多学习实例的模式, 这些模式作为模式匹配算法的模式串。OntoAGS 系统定义的部分模式如下:

模式 1　np1 = ((\tDT \\w +) ? (\tJJ \\w +) ? ((\tNN(S) ? \\w +) +))

模式 2　pnp1 = ((\\tDT \\w +) ? (\tJJ \\w +) ? ((\tNP(S) ? \\w +) +))

模式 3　pnp2 = (pnp1 \tCC(and |or) pnp1)

模式 4　pnp3 = (pnp1((\t, , pnp1) +) \tCC(and |or) pnp1)

模式 5　pnp4 = (pnp1(\t, , pnp1) +)

其中, "\t"代表制表符; "\\w + "代表由一个或多个字符组成的字符串; "? "代表问号左边的内容块可以出现零次或一次; "+ "代表加号左边的内容块可以出现一次或多次; "|"代表从竖线分隔的内容中可以任选一个。

模式 1 的含义是限定词/无 + 形容词/无 + 一个或多个普通名词单数/普通名词复数的组合。

模式 2 的含义是限定词/无 + 形容词/无 + 一个或多个专有名词单数/专有名词复数的组合。

上述模式 1 ~模式 5 是为了定义其他模式而定义的。下面的模式 6 ~模式 8 是实例学习的部分核心模式。

模式 6　pnp_nn = (pnp4 |pnp3 |pnp2 |pnp1) (\t(BEZ |BEDZ |BER |BEDR) \\w +) np1

含义: pnp4/pnp3/pnp2/pnp1 is/was/are/were np1

适用语句举例: Java is a programming language

模式 7　colon = np1 \t: : (np4 |pnp3 |pnp2 |pnp1)

含义: np1: pnp4/pnp3/pnp2/pnp1

适用语句举例: name: Peter

模式 8　such_as = np1 \tDT such \tIn as(pnp4 |pnp3 |pnp2 |pnp1)

含义: np1 such as pnp4/pnp3/pnp2/pnp1

适用语句举例: The language such as English, Japanese and Chinese is very useful

2.4　匹配和还原

把词性标注后的领域文档中的每句话与所定义的模式进行匹配是实例学习算法的重要步骤, 模式匹配过程是能否学习出实例的关键。实例一般都是以某种形式出现的(如复数形式), 所以必须把学习出的实例还原为词干, 否则该实例可能不是所研究领域的正确实例。例 3 详细介绍了实例学习算法的模式匹配和还原过程。

例 3: 使用模式 6 对例 2 中的句子进行模式匹配和还原分析。其中

目标串为　NP MIT BEZ is DT a JJ famous NN university…

模式串为　(pnp4 |pnp3 |pnp2 |pnp1) (\t(BEZ |BEDZ |BER |BEDR) \\w +) np1

模式匹配和还原过程如下:

模式串中包含模式 1 和模式 2, 请参见 2.3 节中对模式 1 和模式 2 的描述。

目标串中的 "NP MIT"符合模式串中的 pnp1, "BEZ is"符合模式串中的(\t(BEZ |BEDZ |BER |BEDR) \\w +), "DT a JJ famous NN university"符合模式串中的 np1, 所以目标串与模式串相匹配。把符合 np1 的 "DT a JJ famous NN university"中 NN 所代表的 university 抽取出, 如果 university 是所研究领域的概念, 则把符合 pnp1 的 "NP MIT"中 NP 所代表的 MIT 从此句话中抽取出并把 MIT 还原为词干, 最后把还原后的 MIT 作为概念 university 的实例。

OntoAGS 系统把实例还原为词干的方法是: 在已有的名词

库中查找此实例,如果名词库中含有此实例,则把实例的原型从名词库中取出,否则不进行任何变换。经调查很多实例是以专有名词单数形式出现,所以 OntoAGS 系统对实例的还原方法可以满足实例还原需要。

3 实验

为了评价 Ontology 系统的实例学习效果,我们把学习实例的数量和准确率作为评价实例学习效果的指标。其中,

$$准确率 = \frac{学习出正确实例的数目}{学习出所有实例的数目} \times 100\% \quad (1)$$

我们从因特网上搜索出大学领域的3 000篇 HTML 文档,并利用这些文档对 OntoAGS 系统和 Text-to-onto 系统的实例学习效果进行了比较。图 1 显示的是在相同的文档数目下,Text-to-onto 系统与 OntoAGS 系统学习出实例数量的比较。图 2 显示的是在相同的文档数目下,Text-to-onto 系统与 OntoAGS 系统学习出实例准确率的比较。

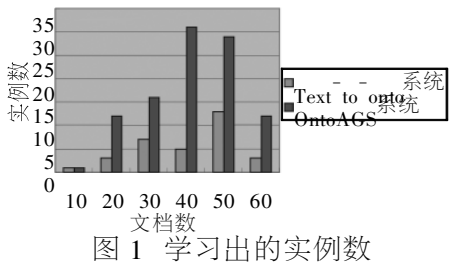

图 1 学习出的实例数

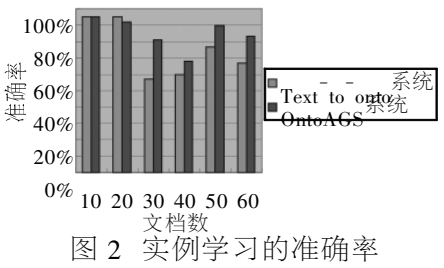

图 2 实例学习的准确率

从图 1、图 2 中可以看出,从相同数目的文档中学习实例,OntoAGS 系统一般都比 Text-to-onto 系统学习出的实例数量多,准确率高。注意:

(1) 从文档中学习出实例的数目并不是随着文档数目的增加而增多,而是与文档中具体的实例数目有关。例如在图 1 中,60 篇文档学习出实例的数目比 40 篇文档学习出实例的数目少,这就是由于实例数量是根据文档的具体内容决定的。

(2) 检验实例学习效果的好坏,既要参考学习出实例的准确率,又要参考学习出实例的数量。从图 2 可知,在对 20 篇文档进行实例学习时,Text-to-onto 系统学习实例的准确率比 OntoAGS 系统学习实例的准确率高。结合图 1 可知在对 20 篇文档进行处理时 Text-to-onto 系统学习出实例的数量比 OntoAGS 系统学习出实例的数量少很多。所以在处理此 20 篇文档时,OntoAGS 系统的实例学习的总体效果是较好的。

4 结论与展望

我们开发了一个 Ontology 自动生成系统 OntoAGS,该系统可以从给定领域的文档集中自动创建该领域的 Ontology,并能自动地从领域文档中学习实例及其与概念的关系。OntoAGS 系统的实例学习是基于模式匹配的算法。实验表明,OntoAGS 系统学习的实例比 text-to-onto 系统学习的实例数量多,准确率高。

我们计划从以下两个方面对实例学习进行深入研究:

(1) 探索模式的自动学习算法。模式的好坏与多少直接决定了基于模式匹配算法的实例学习效果。模式通常是通过观察并人工构造的,这显然不利于获得更多的模式,而且人工的构造错误也会影响实例学习的效果。因此,探索模式的自动学习算法对实例学习的研究是十分有意义的。

(2) 提高词性标注的准确率。由于语言环境的复杂性,词性类别的多样性,运用 QTag 对文档中的词汇进行词性标注的准确率不高。利用 QTag 的标注结果进行实例学习,学习出实例的数目比实际数目少,并且学习出实例的准确率也会降低。所以提高词性标注的准确率能够提高实例学习的效果。

参考文献:

[1] tuder R, Benjamins V R, Fensel D. Knowledge Engineering: Principles and Methods[J]. Data and Knowledge Engineering, 1998, 25(1-2): 161-197.

[2] The Protégé Project[EB/OL]. http://protege.stanford.edu, 2004.

[3] text-to-onto Software[EB/OL]. http://km.aifb.uni-karlsruhe.de/kaon2/download, 2004.

[4] 邓志鸿,唐世渭,张铭,等. Ontology 研究综述[J]. 北京大学学报(自然科学版), 2002, 38(5): 1-2.

[5] OWL Web Ontology Language Guide[EB/OL]. http://www.w3.org/TR/2004/REC-owl-guide-20040210/, 2004-02-10.

[6] QTag Software[EB/OL]. http://web.bham.ac.uk/o.mason/software/tagger/index.html, 1998.

作者简介:

刘贺欢(1981-),女,北京人,硕士研究生,主要研究方向为人工智能;刘椿年(1944-),男,江苏连云港人,教授,博士,研究方向为人工智能。

(上接第 37 页)

3 结论

企业日常应用中经常会遇到多个本体中子本体的抽取问题,本文提出了一种方法,可以方便地从多个本体中抽取所需子本体。此种抽取方法既保证了用户的需求,又降低了抽取的复杂程度,为更深入地研究抽取方法提供了借鉴。统一本体的形式时,术语和关系的表示建议采用 XML 格式,这样有利于制定不同的抽取规则以及抽取复杂的本体。

参考文献:

[1] 王卫东,王英林. 基于企业概念本体的 Web 知识获取[J]. 计算机工程与应用, 2004, 40(16): 191-196.

[2] 张成洪,王向安,古晓洪. 利用 Ontology 和规则表达式的 Web 信息抽取[J]. 计算机工程, 2004, (5): 87-91.

[3] Ricardo de Almeida Falbo, Credinéé Silva de Menezes, *et al*. A Systematic Approach for Building Ontologies[C]. Lecture Notes in Computer Science, Springer-Verlag GMBH, 2003. 349-360.

[4] Wouters C, Dillon T, Rahayu W, *et al*. A Practical Approach to the Derivation of Materialized Ontology View[M]. Web Information Systems, Idea Group Publishing, 2004. 191-226.

[5] S Staab, *et al*. Knowledge Processes and Ontologies[J]. IEEE Intelligent Systems, 2001, 16: 26-34.

作者简介:

刘文斌(1979-),男,山东青岛人,硕士,研究方向为 MIS 系统与分布式数据库、企业知识管理等;谢强(1972-),男,四川绵阳人,博士,研究方向为企业信息化、分布式数据库等;张磊(1977-),男,江苏徐州人,博士,研究方向为企业信息化、知识物流等。