计算机科学 2005Vol. 32№. 11

基于Ontology的网页关系识别研究

樊闻斌　丁　艳　潘金贵
（南京大学软件新技术国家重点实验室　南京大学计算机科学与技术系　南京 210093）

摘　要　对万维网上大量的网页之间进行特定语义关系的分析，将这样的工作用于搜索引擎中，可以实现智能化的查询，和提供其它个性化服务。本文借助于Ontology中的关系实例，在网页分类的基础上对网页之间的关系进行自动识别，同时提出了网页关系识别规则的自动生成和优化方法。

关键词　Ontology，网页关系识别，万维网内容挖掘

Recognition of Web Page Relation Based on Ontology

FAN Wen-Bin　DING Yan　PAN Jin-Gui
(National Key Laboratory of New Software Technology, Computer Science of Nanjing University, Nanjing 210093)

Abstract　This paper analyzes semantic relation among large amount of web pages in WWW, which can realize intelligent query and provide characteristic services when applied to search engine. When recognizing the relation of web pages, this paper realizes the automatic production of rules through relation entity and sample sets defined in Ontology and recognition based on rules. Based on the idea of FOIL and the method of SRV, we reduce the training process of rules when training rules through sample sets. Through regulating and comparing the valve of characteristic selection, we can get the optimization rule sets and simplify the process of judgment when precision is improved.

Keywords　Ontology, Recognition of web page relation, Web content mining

1　引言

自从1990年出现了世界上第一个真正意义上的网页[1]后，网页数的增长呈指数形式[2]，每过不到9个月的时间网页总数就会翻一番[3]。万维网的快速发展，构成了一片信息的海洋，人们庆幸摆脱了信息匮乏的困境之余，又因为万维网的海量性、复杂性、分布性以及非结构性而陷入了“信息过量”的境地。

面对这样的问题，人们一直在探寻各种各样的解决方法。这些解决方法大致可以分为两个途径：一是从万维网的信息表示出发，通过研究和制定有效的万维网表示规范来促进信息的规范表示，从而方便人们在此基础上开发有效的信息检索等服务。如开发XML语言让计算机可以识别和自动处理网页文档包含的信息，推广资源描述框架(Resource Description Framework)，以便提供一个通用的万维网资源描述规范，实现知识表示和共享的Ontology技术，等等。二是可以从研究万维网的信息特性出发，通过开发特有的搜索技术和分析处理技术来为人们提供有针对性的、高效的信息获取服务。例如，人们熟知的搜索引擎，万维网信息挖掘等。

以上的解决方法和技术都各有优缺点。我们将上述两种途径结合起来，在Dolphin* 系统中，把信息表示中的新兴技术——Ontology与万维网内容挖掘(Web Content Mining)相结合，利用了Ontology中的关系实例，在Dolphin系统已经完成分类的基础上[4]利用规则对网页之间的关系进行了自动识别。我们借鉴了一些机器学习的方法，根据主题资源的特点，提出了实现规则的自动生成和优化的方法。最后将方法应用到了系统中，获得了较好的效果。

2　相关工作

目前网页关系识别方面的研究成果很多，使用的主要方法还是机器学习，具体的实现则各不相同。对关系的学习也可称为归约逻辑编程(Inductive Logic Programming, ILP)，主要通过规则学习的方法来生成一系列的规则，进而对其他具有类标号的实例进行关系识别。这些研究的基础是Web的“二元性”[5]，它能够将两组有内在联系的概念标识出来。

对关系的学习方法通常基于一阶规划学习算法(FOIL)[6]，用谓词的表示形式组成规则。假设要从Web上将有关论文的一个(Author, Title)关系提取出来，可以产生如下一条规则：

Author_of (A, B):- Person(A), Publication(B), link_to(A,B)

其中，A和B代表Web文档，Person(A)和Publication(B)代表该文档的类别谓词，link_to(A,B)检测是否有从A到B的超链，以上规则是基于Horn子句形式的。

FOIL是一个从实例构建Horn子句规则的学习系统，尤其能够识别各项之间的关系。它的输入是有关这些关系的信息，其中之一(目标关系)被定义为Horn子句形式。对每个关系都有一个属于这个关系的常量元组集合。对于目标元组

樊闻斌　硕士研究生，主要研究方向为Web信息检索；丁　艳　硕士研究生，主要研究方向为Web信息检索；潘金贵　教授，博士生导师，从事中间件、Agent技术及多媒体远程教学系统等的研究。

* Dolphin系统是由南京大学多媒体研究所远程教育实验室开发的课件搜集系统，其前身是DOLTRI-Agent系统(Distance and Open Learning Training Resource Information retrieval Agent)，它的中文名称是基于RDF和关键字的专用智能网络信息搜集代理(Intelligent Web Information Retrieval Agent)。

还会有一些反例。FOIL 的学习过程从子句的左边开始，通过往右边添加谓词项不断地深入，直到没有反例被这个子句覆盖到或者子句过于复杂(超长)为止[7,8]。

在 CMU 的 Web→KB 系统中，采用了一种类似于 FOIL 的方法，但做了改变[9]，当为某个概念添加子句时不采用 FOIL 中那种简单的"爬山式"搜索，并且搜索过程采用了不同的评估函数。这种搜索过程分为两个阶段，首先，对子句的"路径"部分进行学习，然后使用 FOIL 方法为子句添加条件项。对子句的路径部分采用了 Richards 和 Mooney 的"关系规则发现"方法获取规则的 path 部分，然后自下而上的搜索增加短语，直到这个规则只覆盖或者大部分覆盖所有的正例。这种方法可以识别间接关系。

文[10]使用了 SRV 方法进行关系挖掘，通过特征文本的提取和关系路径的识别获取 Web 文档关系。SRV 的优点在于效率比 FOIL 高一些，并增加了验证的过程，特别适用于 Web 文档的信息抽取。

由于 Dolphin 系统采用的 Ontology 中已经有了许多对领域内各个概念之间的关系描述(如 Employed. By，Produced. By 等)，巧妙地利用这些关系的定义，并采用合适的方法，训练出一个规则集，便可以高效地识别出网页之间语义上的联系。在规则集的生成上，FOIL 方法有个比较突出的缺点就是效率不高，因为其基于了样本遍历的思想。因此，我们吸收了其中的一些思想，借鉴了 SRV 的过程，采用了一些自定义的、高效快速的特征提取和判别方法来生成规则。

3 基于 Ontology 的网页关系规则训练机制

Web 上的网页不是互相孤立的，它们之间不仅在 HTML 语法上存在着联系，而且在语义上也是紧密相连的。如一个教授的主页，可以指向他的课程、他的著作、他的工作单位等其他信息，这些东西如果单凭普通的关键字搜索是很难描述的，但是如果可以识别出这些网页之间的关系，那么只需搜索这个教授的名字便可以定位到相关的其他页面，这样不管是准确率还是效率都大大提高。我们接下来所要做的，就是在获取网页的类别信息基础上，依靠特点方法自动地识别出这些已分好类别的 Web 文档之间的关系，为智能化的语义搜索提供有力的支持。

3.1 Ontology 概述

Ontology 就是关于某一领域对象的类别、属性及它们之间的关系的内容理论(Content Theory)[11]。在计算机学科中，我们狭义地把它看成是关于某一领域的概念性描述。一个典型的 Ontology 包含领域中的层次化的概念体系，并通过属性-值这一机制来刻画其中的概念。从某种意义上讲，它是一种领域性的知识表示。

Ontology 以其对知识的概括和描述能力在 Web 上得到广泛的应用。一般 Web 上的 Ontology 包括分类和一套推理规则。分类定义了对象的类别及其之间的关系，并且是一个可继承的层次结构，这与面向对象思想中的类、子类，实体间的概念是一致的[11]。通过给类指定属性，允许子类继承类的属性，这样我们就能够表达实体之间的大量关系；Ontology 中的推理规则提供进一步的功能。正是由于 Ontology 本身的这些特点，可以运用它增强网络服务的功能。辅以简单的方法，它们就能改进网上搜索的准确性，使搜索程序只寻找那些指向精确概念的网页，而不是仅仅通过模糊关键字查到的所有页面。

在我们的项目中，初步使用的 Ontology 主要参考了 SWRC[12](Semantic Web Research Community Ontology)，并进行了一些裁减。使用该 Ontology 是因为自己构造 Ontology 是一项浩大而艰苦的工作，需要具备许多本领域的相关知识，目前 Ontology 的构造一直是研究的热点。SWRC 已经经过了若干版本的修改，它能够很好地描述 Web 网页的性质。在我们的项目中，搜索的信息主要偏重于科研方面，SWRC 这个 Ontology 正好提供了该方面的一个描述。

3.2 识别 Ontology 中的关系定义

```
〈daml:Class rdf:ID="AcademicStaff"〉
    …
    〈rdfs:subClass Of〉
        〈daml:Restriction〉
            〈daml:onProperty rdf:resource="#headOf"/〉
            〈daml:toClass rdf:resource="#Project"/〉
        〈/daml:Restriction〉
    〈/rdfs:subClassOf〉
    …
〈/daml:Class〉
```

图 1 Ontology 中的关系定义(一)

在 Dolphin 系统的 Ontology 中，不仅定义了领域内的主要概念，还定义了各个类别之间的关系，参见图 1 给出的例子。在 AcademicStaff 类中，定义了该类和 Project 等类的关系，它与 Project 是"headof"关系，即某个教工如果和某个项目联系起来，那么很可能是该项目的 head。借助这种定义，对于有着超链关系的两个页面，只要正确识别出它们所属的类别，那么便可以很快地根据 Ontology 定义找到它们之间可能存在的逻辑关系。对于上例，我们可以定义如下规则来识别这些关系：

headof(A,B):- AcademicStaff(A)，Project(B)，link_to(H,A,B)

其中，link_to(H,A,B)表示网页 A 中存在指向 B 的链接 H。

但是，并不是所有情况都像上面那么简单。参考图 2 给出的另一个例子，也是系统中 Ontology 的一部分摘录。在这段定义中，描述了"Project"类和"Organization"类之间的关系，一是"carriedOutBy"，另一个是"financedBy"，即如果某个项目和某个组织的页面有联系，那么该组织可能是这个项目的执行者，也可能是赞助者，在这种情况下，该如何区分它们呢？所以说，各个类别之间的关系不是唯一的，因此若要正确地识别它们，要使用一些方法识别这两种关系之间不同的特征，确定更多的区分条件并加入规则中。这就依赖于高效的方法。

```
〈daml:Class rdf:ID="Project"〉
    …
    〈rdfs:subClass Of〉
        〈daml:Restriction〉
            〈daml:onProperty rdf:resource="#carriedOutBy"/〉
            〈daml:toClass rdf:resource="#Organization"/〉
        〈/daml:Restriction〉
    〈/rdfs:subClassOf〉
    〈/rdfs:subClassOf〉
        〈daml:Restriction〉
            〈daml:onProperty rdf:resource="#carriedOutBy"/〉
            〈daml:toClass rdf:resource="#Organization"/〉
        〈/daml:Restriction〉
    〈/rdfs:subClassOf〉
    …
〈/daml:Class〉
```

图 2 Ontology 中的关系定义(二)

3.3 网页关系识别过程及特征标注

关系存在于两个相关的 Ontology 类别中，也可以看成存在于两个网页(簇)之间，这两个网页(簇)通过链接发生联系。一旦网页的 Ontology 类别已经被识别出来，通过分析联系网页的链接，及描述链接的文字，匹配规则集中的规则，就可以确认网页间具体的关系。具体的作法概括如下：

(1)识别网页 a 的类别；

(2)识别网页 a 所指向的网页 b 的类别；

(3)抽取链接信息，检查规则集；

(4)应用规则，确定关系。

在这里，规则集是预先根据样本分析得来的，它告诉我们，链接信息在具有什么样的特征下，网页间具有什么关系。链接信息的特征主要靠周围的文本确定，因此可以靠对超链周围的文本进行分析而生成完整的特征。借鉴了 Web→KB 中的表示方法[9]，我们主要通过如下特征来描述关系：

(1)出现某个单词：has_word(H,W)，即锚文本 H 中出现了某个特征单词 W，而这个特征单词是通过样本训练出来的属于该关系的特征词典中的一个。如My Publication。

(2)出现包含数字的单词：has_numeric_word(H)，即锚文本 H 中出现了包含数字的单词，如 I am a member of 97CS。

(3)周围文本出现单词：has_neighbor_word(H,W)，即锚文本 H 附近的地方出现了词典中的单词 W。如 carried out by Dolphin Group。

(4)出现首字母全部大写的单词：all_words_capitalized(H)，即锚文本 H 中的单词的首字母全部是大写的，如 financed out by Nanjing University。

就锚文本的主要结构和它所包含的信息量来看，以上四类特征基本能标识和刻画出系统 Ontology 里所描述的各个关系。通过这些特征，就可以较为准确地定位到两个类别的 Web 文档之间的关系。

3.4 规则集的产生

我们的规则抽象为如下形式：

R (A, B):- Class1(A),Class2(B),[Condition]*

其中，A 和 B 的类别是已知的，主要的工作就是集中在提取其他特征作为规则成立的辅助条件，提高准确率。我们首先对 Ontology 中的关系描述进行提取，形成初始规则集。然后以此为基础，按照如下方法对这些初始规则进行扩展：

(1)设对应于该类关系的样本集为 T，T 中的元素是关系实例，以 R(A,B)的形式表示，R 为 Ontology 中描述的关系名称；

(2)对于每个样本集中的样本元组 t(A,B)，如果 link_to(H,A,B)成立，则提取出 A 中指向 B 的锚文本放入锚文本集合 S_1 中，该文本周围 n 个字符内的近邻文本放入近邻文本集合 S_2 中；

(3)分析 S_1，用文本分类中提取特征向量的提取方法获取对应于该类关系的锚文本特征字典 D_1；由 S_2 分析获取近邻文本特征字典 D_2；

(4)对于任一 $s \in S_1$，顺序进行如下过程：

如果 s 中单词全部大写，则添加条件 all_words_capitalized(H)，计数器加 1；

如果 s 中出现包含数字的单词，则添加条件 has_numeric_word(H)，计数器加 1；

(5)将所有计数与样本数相比得到支持度，对于 D_1 中支持度大于某一阈值 α 的单词 w，添加条件 has_word(A,w)到规则前件中；对于 D_2 中支持度大于 β 的单词 w′，添加条件 has_neighbor_word(A, w′)到规则前件中；将(4)中统计得出的阈值大于 γ 的对应条件添加到规则中。α、β、γ 的选取对于规则的覆盖率和准确率有着很重要的影响，下一节的实验结果将对此做出比较。

(6)检查规则，对于前件中的 Class 条件相同者，即 A 与 B 中可能存在多种关系的情况，检查其他前件，若有完全一样的条件，则作为多余条件删除。这个步骤是对规则的一个简化，可以加快规则的判断速度。

在图 2 的情况中，对于 Project(A)，Organization(B)，按照以上方法获取的规则可能有如下形式：

carriedOutBy(A, B):- Project(A), Organizaiton(B), has_neighbor_word(H,"carry"),link_to(H,A,B)

financedBy(A,B) :- Project(A),Organizaiton(B), has_neighbor_word(H, "finance"),link_to(H,A,B)

总的来说，上述方法不仅在规则的训练方法上进行了简化，而且在规则的表述方面也做了一些消除冗余的工作。在寻求准确高效的同时，力求简化，这也是基于 Ontology 的网页关系识别的特点所在。

上述方法只是识别出了直接关系，是建立在 R(A,B)中的 A、B 间有直接超链关系的基础上，如果 link_to(A,B)不成立，即 A 与 B 之间是有间接关系的，这种情况该如何处理？请参看图 3 给出的例子。

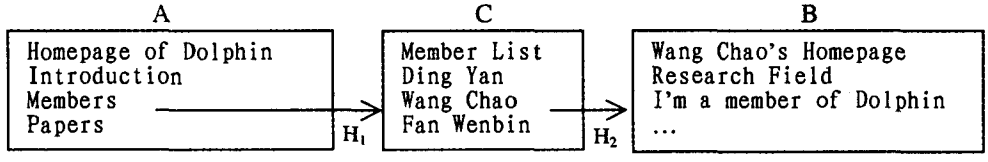

图 3 具有间接关系的相关网页示例

网页 A 属于 Project 类别，B 属于 Person 类别，A 和 B 之间有着 members_of_project(A,B)关系，但是 A 与 B 没有直接的超链关系，因为 A 通过一个链接 H_1 指向一个 Member-List C 来描述它的成员，只有 C 中才有指向 B 的链接 H_2。但是，样本库中只有 members_of_project(A,B)这个关系实例，如何获取 A 与 B 之间的规则呢？

可以看到的是，其他的基本条件如 H_1 的特征，A、B 的特征是可以不变的，关键是找到由 A 到 B 的路径。这实际上就是对超链的拓扑结构图进行搜索，找到存在于两个页面 A 和 B 中的一条或多条路径的搜索过程。具体搜索过程我们采用了深度优先的算法，并且对搜索深度进行了阈值限制，主要是因为深度优先可以节省时间，比较快速地找到目标路径，当路径过长时还可以退出。

这样，在找到从 A 到 B 的路径后，只要在前件中添加有关路径的条件子句，即可形成完整的规则。如上例所示的 A->C->B，可以获取如下规则：

members_of_project(A,B):- Project(A),Person(B),link_to(H_1,A,C),

link_to(H_2,C,B),has_word(H_1,"members")

4 实验结果分析

我们根据Ontology中的定义,选择了20个关系在包含400个关系实例的样本集中进行规则的训练,并对这些规则的正确率进行了检验。从中选取了3条规则如下:

(1)employs(A,B):- Organization(A),Person(B),link_to(L_1,A,C),link_to(L_2,C,B),has_neighbor_word(L_1,"member")

(2)editor(A,B):- Person(A),Publication(B),link_to(H,A,B),all_words_capitalized(H)

(3)publishes(A,B):- Organization(A),Publication(B),link_to(H_1,A,C),link_to(H_2,C,B),has_neighbor_word(H1,"publication")

如前文所述,α、β、γ的选取对于规则前件数量的影响是比较大的,有时候还能产生多个规则。从我们的方法中不难看出,我们考虑的不是单个样本文档的情况的综合,而是针对特征条件的统计规律进行规则的生成。这种作法的好处在于能够尽可能多的获取一些通用的特征作为规则前件,规则的准确率较高,缺点是训练出来的规则进行匹配时可能覆盖率较低。因此,我们规定,上述的几条规则都是在$\alpha=\beta=\gamma=0.8$的情况下训练出来的。对于规则(1),如果设定$\gamma=0.7$,则会增加一个前件all_words_capitalized(L_2),成为如下形式:

employs(A,B):- Organization(A),Person(B),link_to(L_1,A,C),link_to(L_2,C,B),

has_neighbor_word(L_1,"member"),all_words_capitalized(L_2)

而(2)则在$\alpha=0.7$的情况下变为

editor(A,B):- Person(A),Publication(B),link_to(H,A,B),

all_words_capitalized(H),has_word(H,"publication")

相应地,识别的准确率和覆盖率也会随着α、β、γ的值而变化。通过改变这些阈值,我们获取了上述三条规则的覆盖率与准确率之间的关系变化如图4所示。从结果中可见,规则1的总体效果最好,这表示定义的特征条件对其具有显著效果,规则3表现较差,原因在于其识别的为间接关系,前件较多,其覆盖率也较低。总体来说,α、β、γ的值不能设定的太高或者太低,维持在0.7~0.8的范围内比较适中。其中α的值不能过低,否则会带入更多的无关条件,使得覆盖率较低给规则的判断增加了难度;而β、γ的过高也会使得条件的选取不够而导致规则过于简单,从而准确率降低。

此外,分类的准确率对于规则的准确率的影响也较大。因为某个网页属于哪个类别对于关系识别规则来说是必要条件,分类的成功与否也和关系识别的准确率有着很密切的关系。当然,我们识别出来的规则是在比较好的样本基础上的,样本集和测试集里所有的实例是在进行了正确分类的基础上训练的。因此,实际的规则识别正确率比上表所示要更低一些。但是,总体来看,因为充分发掘了结构信息和文本信息的潜在语义关系,识别的效果还是不错的。

总结及进一步工作 本文讨论了基于Ontology的网页关系识别。进行网页的关系识别时,我们采用基于规则的方法进行识别,并且借助于Ontology中定义的关系实例和样本集实现了规则前件的自动生成。在通过样本集训练规则时,我们对一些前件特征进行定义,并且在FOIL思想和SRV方法的基础上简化了规则的训练过程,通过调整特征选取的阈值进行比较以求获取最优的规则集,使准确率能够得到提高的同时判断的过程也得到简化。我们将这一网页关系识别的过程应用到Dolphin系统中,使用户能够在系统的指导下,快速的定位到与起始页面有一定语义关系的其他类型页面中,大大节省了浏览和寻找的开销。

但是这种网页关系识别目前的应用范围还较窄,在我们的实验中,仅限于系统选择的几个类别之间的关系识别,且目前就我们自己使用的Ontology定义来看,其关系实例的定义还不够全面。但是,只要能够提供通用的接口来获取这些规范信息,这种关系的识别将很容易地扩展到其他领域和范围。

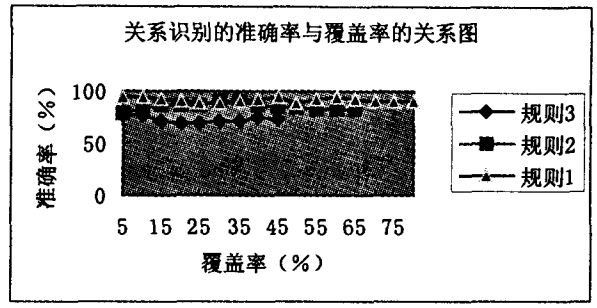

图4 网页关系识别规则的效果

此外,关系识别不仅仅限于某两个或几个网页之间,还包括存在于单个页面内部及页面组之间的关系识别,这些则需要进行更多的研究和实践,采用一些新的方法。目前采用的关系识别方法因为强调效率而忽略了规则的验证调整过程,这也是一个值得完善的地方。这些都将是以后这部分工作的前进方向。

参考文献

1 Connolly D. A Little History of the World Wide Web. http://www.w3.org/History.html, 2002.11

2 Barfourosh A A, Anderson M L. Information Retrieval on the World Wide Web and Active Logic: A Survey and Problem Definition, http://www.cs.umd.edu/Library/TRs/CS-TR-4291/CS-TR-4291.pdf, 2002

3 Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. Addison Wesley, 1999

4 丁艳,曹倩,王超,潘金贵. 基于Ontology和EM方法的网页分类研究. 计算机科学,2003,30(11):112~115

5 Sundaresan N, Yi J. Mining the Web for Relations. Computer Networks: The International Journal of Computer and Telecommunications Networking. Amsterdam, Netherlands: North-Holland Publishing Co, 2000. 699~711

6 Mitchell T M. 曾华军,张银奎,等译. 机器学习. 北京:机械工业出版社,2003. 204~208

7 Quinlan J R, Cameron-Jones R M. FOIL: A Midterm Report. In: Proc. of the European Conference on Machining Learning. Vienna, Austria: Springer Verlag, 1993. 3~20

8 Quinlan J R. Learning Logic Definitions from Relations. Machining Learning, 1999, 5:239~266

9 Craven M, DiPasquo D, Freitag D, et al. Learning to Extract Symbolic Knowledge from the World Wide Web. In: Proc. of the 15th National Conf. on Artificial Intelligence (AAAI-98). Madison, US: AAAI Press, 1998. 509~516

10 Freitag D. Information Extraction from HTML: Application of a General Machine Learning Approach. American Association for Artificial Intelligence. Madison, US: AAAI Press, 1998. 517~523

11 Chandrasokaran B, Josephson J R. What Are Ontologies, and Why Do we Need Them? IEEE Intelligent Systems, 1999, 14:20~26

12 Sure Y. Semantic Web Research Community Ontology. http://ontobroker.semanticweb.org/ontos/swrc.html, 2001-12-11