面向半结构化文本的领域本体关系抽取*

程晓[1] 郑德权[1] 杨宇航[1] 邵国军[2]
[1]哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001
[2]秦皇岛首信自动化系统工程有限公司 秦皇岛 066004
E-mail: chengxiao995@yahoo.com.cn

摘 要：本文提出了一种以半结构化文本作为数据源，进行领域本体关系抽取的方法。首先，利用概念实例和属性值的共现得到文档集合。其次，定义关系模式形式，从文档集合中得到关系模式实例，包括关系模式实例的聚类以及类内合并。最后，将各类关系模式用于抽取领域本体新实例的属性值信息。在针对电影，图书和音乐三个领域进行的实验中，关系模式聚类的错分率和漏分率分别为 0.19%，1.31%，类内合并后关系模式的准确率最高可达 85%。实验结果表明了本方法对于领域本体中关系抽取的有效性。

关键词：本结构化，领域本体，关系抽取，模式实例

Automatic Domain-Ontology Relation Extraction from Semi-Structured Texts

Cheng Xiao[1], Zheng Dequan[1], Yang Yuhang[1], Shao Guojun[2]
[1]Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001
[2]Beijing Shougang Automatic Information Technology Co., Ltd, Qinhuangdao 066004
E-mail: chengxiao995@yahoo.com.cn

Abstract: This paper presents a new method to acquire Domain-Ontology relations from semi-structured data sources. First, obtain documents according to the co-occurrence of concept instance and attribute value. Further, define formats of relation patterns, and extract pattern instances from documents, including pattern clustering and pattern combining in each cluster. Finally, relation pattern instances are applied to gain attribute values of new concept instances in Domain-Ontology. Experiments are carried out in fields of film, book and music, the rate of pattern incorrect-division and pattern leakage are respectively 0.19% and 1.31%, the highest precision of combined relation patterns reaches 85%. Experimental results demonstrate that the method developed in this paper is fairly efficient.

Keywords: semi-structure, Domain-Ontology, relation extraction, pattern instance.

1 前言

随着互联网的飞速发展，信息量极大丰富，快速构建和更新领域知识库的需求更加迫切。包含特定领域概念的术语是任何领域描述概念的最基本单元，然而仅仅有术语还不足以充分的描述领域知识，更重要的是找到术语间的关系并将其定位到现存本体的适当位置。领域本体包括领域术语、概念、概念和术语间以及概念与概念间的各种关系[1]。因此领域本体能够表达丰富的领域知识并能有效的应用于各种自然语言处理和知识工程的任务中[2]。

网络资源日益丰富，几乎包含所需的各个领域的信息，如何能够从这一大规模的半结构化文档集合中得到各种结构化的信息成为了当前研究的热点和重点。本文以半结构化的 Web 网页作为数据源，从中自动生成领域本体的关系抽取模式，并且将关系模式应用于领域本体的自动扩展和更新中。

*本文承国家自然科学基金（项目号 G60736044）和国家 863 计划重点项目(项目号 G2006AA010108)的资助。

2 相关研究

关系抽取按照其使用特征信息的不同分为直接使用上下文信息、间接使用上下文信息和使用内部信息的关系抽取。按抽取方法可分为基于分类器[3]和基于模版[4]的关系抽取。

直接上下文信息指直接描述或嵌入关系的上下文。利用此种特征的方法包括模版抽取、关联规则挖掘、分类以及统计度量等。基于模版的抽取方法[5]召回率普遍较低，为了解决这样的问题，PANKOW[6]和C-PANKOW系统[7]使用Google等搜索引擎进行关系抽取。ACE举行的关于实体间关系检测和抽取的比赛中，以SVM为代表的分类算法[8]在评估中取得了良好的效果。

间接上下文信息指并不直接包含关系，却可以通过比较不同概念间的上下文信息推出关系。使用上下文搭配的层次聚类算法[9]可以抽取一般/特殊等关系。形式概念分析是Wille提出的一种从形式背景建立概念格来进行数据分析和规则提取的强有力工具[10]，也是利用间接上下文信息进行关系抽取的典型方法[11]，FCA是数据分析和知识表达的形式工具，FCA使用两个数据集，一个对象集和一个属性集，找到两个数据集之间的二元关系，进一步的构造被称作形式概念格和根据形式概念的概念包含序列。

语义解释是典型的利用内部信息进行关系抽取的方法[12]，通过决定复杂术语每个组件的正确概念，识别概念组件间的相互关系，从而构建复杂概念的过程。

3 方法描述

在本文中，以半结构化的Web网页作为数据源，利用实体间的共现从中进行领域本体的关系模式抽取，进而将关系模式应用于领域本体中其它实例的信息抽取，实现领域本体的扩展。

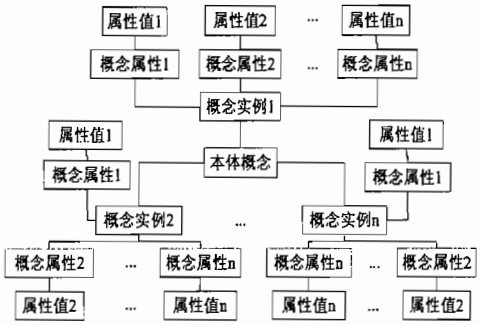

图 1 领域本体组织结构

本文中定义的本体结构如图1所示。本体概念可以实例化为若干个概念实例，本体概念可以用若干个属性描述，由于本体概念是对本体中实例的抽象，每个概念实例也用相同的属性来描述，针对于不同概念实例，不同的属性有唯一的属性值与之匹配。例如，电影领域中，本体概念为电影，属性为导演，主演等，概念实例可以包括长江七号，大灌篮等电影实例，（导演，周星驰）就是长江七号的属性值对。在本文中主要研究的是实例和各个属性值之间的关系。

3.1 领域本体关系抽取

本文中的领域本体关系抽取方法首先获得领域文档集，定义关系模式，并利用概念实例和属性值的共现抽取关系模式实例，经过去重，聚类和合并操作之后得到若干类的关系模式实例。具体的抽取步骤如下所示：

Step1：采样网页

将实例名和属性值作为 query，利用 Google 搜索引擎得到 Web 网页片段。

Step2：获得领域文档集合及文档预处理

从文档中选择实例和属性值共现的行。分词并修正，分别将实例名和属性值作为一个整体。

Step3：抽取关系模式实例

逐行定位实例和属性值，将其出现的上下文提取出来。本文中将模式形式定义为：$H_1SH_2XH_3$，或者$H_1XH_2SH_3$。其中S为实例，X为属性值，H_1为实例S的前缀词，H_3为属性值X的后缀词，H_2为实例和属性值之间的词串。

Step4：过滤关系模式实例

若得到的模式中实例和属性之间的字符串H_2的长度大于某个阈值上限，或者H_2中含有某些停止符，或者含有实例名，人名，则过滤掉该模式。

Step5：相同模式实例的去重和相似模式实例的合并。

首先，为了判断两个模式是否相似，定义一个相似度函数$SimilarDegree(P_i,P_j)$，

$$SimilarDegree(P_i,P_j)=\frac{ComWords(P_i,P_j)*weight_1+ComPair(P_i,P_j)*weight_2}{Maxlen(P_i,P_j)}$$

其中，P_i，P_j分别代表两个模式，$ComWords(P_i,P_j)$表示P_i，P_j中公共词汇串的长度；$ComPair(P_i,P_j)$表示P_i，P_j中连续公共词汇的对数；$Maxlen(P_i,P_j)$表示模式P_i和P_j中长度的最大值；$weight_1$，$weight_2$表示权重，并满足$weight_1+weight_2=1$。

其次，进行聚类和类内合并，每一类最终得到若干个合并后的模式。

在聚类时，采用单链法聚类。模式合并就是不断将同一类别的模式集中两个相似度最大的模式合并，直到最大相似度小于某个阈值时停止合并，得到若干抽取模式。

经过模式的聚类和类内的相似合并，得到若干组关系模式实例，每一组模式代表一类关系。

3.2 领域本体扩展

在前面我们已经得到了若干类实例和属性值之间的关系模式，在构建领域本体时，需要对其进行实时更新，将新实例的属性值的信息加入本体结构中，从而使得领域知识更加丰富。

Step1：获得新实例的 Web 文档

Step2：利用关系模式抽取新实例的属性值信息

遍历各 Web 文档，如果某文本片段能够匹配关系模式$H_1SH_2XH_3$，或者$H_1XH_2SH_3$，其中S代表新实例名称，X 就是所需的属性值信息。对于每个文本片段，分别使用前向和反向的

模式匹配方法获得。

Step3：将新实例的属性值对信息以（属性，属性值）二元组的形式表示出来

经过以上步骤，领域本体可以得到不断的实时扩展，从而形成领域知识库，为各种自然语言处理任务提供帮助。

4 实验结果及分析

本文在电影，图书，音乐三个领域进行实验，对于每个领域，需要的输入信息包括领域属性，领域实例以及实例的属性值对。输入信息是本课题有关领域概念属性和属性值对抽取的工作中获得的结果，选择最常用的属性和各个属性值对用于本文实验，抽取过程不再详述。本实验采用十个领域实例，分别与属性值组成 query，利用爬虫为每个实例-属性值对获得 1000 个网页片段。

本文中相似度函数公式中两个权重系数 $weight_1$， $weight_2$，以及相似度阈值 T，设 $D=\frac{weight_1}{weight_2}$，下面两组实验以电影领域为例分别验证 D 和 T 对关系模式抽取的影响。

表1 T 恒定，D 变化的模式聚类结果

T=0.5	错分率	漏分率
D=0.7	0.38%	1.43%
D=1.0	0.25%	1.40%
D=1.5	0.19%	1.46%

在表 2 中，相似度阈值 T 保持不变，可以看出，权重系数的比值 D 越高，模式的错分率越低，而漏分率相差不大。这说明，在判断模式之间相似度时，两个模式之间的公共词汇对数这个特征相比于公共词汇长度这个特征对于相似度的判别贡献更大。

表2 D 恒定，T 变化的模式聚类结果

D=1.5	错分率	漏分率
T=0.4	0.28%	1.31%
T=0.5	0.19%	1.46%
T=0.6	0.24%	1.53%

从表3中可以看出，当权重系数的比值D不变时，相似度阈值T越高，模式的漏分率越高。这是由于一些模式实例的长度较长,而相互间公共子序列的长度又比较短,导致相似度值很低,从而被错误地分开。

经过聚类和类内合并之后，对于每种关系类型均得到若干组不同的模式实例，在此，为了判断各类模式实例的有效性，将各种关系模式用于抽取10个领域新实例的属性值信息，根据属性值的正确性来判断模式的有效性和准确性。

从表3中可以看出，不同类型的关系模式，其准确率也有一定的差异。其中“主演”这种关系类型的准确率达到85%，原因主要有：第一，该类信息分布有很强的规律性，因而关系模式的代表性很强，适用性也相应的很广；第二，该类信息一般是用户比较关心的信息，语料也很容易获得，且质量相对较高。某些类型的关系模式的准确率相对较低，例如“片长”，原因有以下几点：第一，模式聚类不合理导致关系模式的通用性不强，无法用于其它实例信息的抽取；第二，

有关新实例的语料中没有相应的属性值信息，使用关系模式获得的都是错误的信息；第三，新实例语料中实例和属性值的分布形式多样，关系模式相对较少。

表3 电影领域关系模式准确率分析

关系类型	类别数	模式实例个数（合并后）	正确模式数	准确率
导演	27	43	36	83.7%
主演	47	80	68	85.0%
别名	8	18	14	77.8%
国家地区	16	29	22	75.9%
片长	10	25	18	72.0%
年代	24	35	28	80.0%
类型	41	57	48	84.2%
语言	15	18	15	83.3%
编剧	32	48	38	79.1%
出品	19	30	23	76.7%

表4 图书领域关系模式准确率分析

关系类型	类别数	模式实例个数	正确模式数	准确率
作者	11	20	15	75.0%
出版社	18	42	30	71.4%
出版日期	24	52	41	78.8%
所属分类	13	21	14	66.7%
ISBN	20	32	22	68.7%
定价	12	23	18	78.2%
开本	9	24	17	70.8%
版次	15	13	9	69.2%
页数	19	21	13	61.9%
字数	23	28	19	67.8%

表5 音乐领域关系模式准确率分析

关系类型	类别数	模式实例个数	正确模式数	准确率
歌手名称	29	55	46	83.6%
专辑语种	18	34	24	70.5%
唱片公司	10	21	14	66.7%
发行时间	14	25	19	76.0%
专辑类型	19	23	14	60.9%
地区	21	39	25	64.1%

三个领域的关系模式准确率均较好，但从表4和表5中可以看出，在图书和音乐这两个领域中关系模式的准确率相比于电影领域较低，原因主要归为以下几点：第一，电影作为一个比较大的领域，能获得的各方面的描述信息也很多，同时属性值对的结构特征对于整个页面变化很小。而图书和音乐领域从网上能获得的以文本形式存在的领域相关文档较少，导致关系模式的特殊性太强，通用性太弱；第二，语料中信息不全，只提到属性，并没有涉及属性值，例如图书领域的“页数”，音乐领域的“专辑类型”，即使正确的关系模式也不能获得相关属性值，导致准确率较低。

本文中的关系抽取可以看作IE（Information Extraction）中的一个方面，但又不完全等同于IE

中的关系抽取。传统IE中的关系抽取是从文本中获得用户所关注的某种类型的信息[13]，而信息的类型是预先指定的，例如公司和产品的关系，本文中的关系抽取并不限制关系类型，而是将文本中出现的所有类型的关系都抽取出来。另外，IE中使用的文本一般都是格式统一的，例如，浙江大学图书馆主页上的“浙江大学被SCI收录的论文检索”中2001年度的所有数据，作为实验对象抽取论文信息。这些文本不论从格式还是组织上几乎是完全一致的，因此通过规则进行信息抽取即可得到很到的准确率，一般可达到90%左右。而本文中的关系抽取，准确率最高为85%，相比于传统的信息抽取较低的原因有：首先，采用的领域文本是利用Google搜索引擎随机获得的Web网页，因为来自于不同的网站，文本之间并不存在完全一致的信息结构特征；其次，本文中获得大量模式实例之后进行了模式的聚类和类内合并，也会有一定的误差。

5 结束语

本文提出了一种利用 Web 网页文档作为数据源，抽取领域本体中概念实例和属性值之间关系的方法。通过文档片段中的共现特征得到两组不同形式的关系模式实例，经过模式实例的聚类和合并来获得不同类别的关系模式，进而抽取领域本体中新实例的各种属性值信息。在以后的工作中也存在几点仍需改进的地方：第一，关系抽取模式的生成是一个渐进的过程，模式实例聚类算法应进一步完善以适应不断增加的相关文本；第二，模式合并过程中，对两个模式实例的合并操作过于简单，需要增加语义属性等信息以提高合并的正确性。

参 考 文 献

[1] DU Xiao-Yong, LI Man1, WANG Shan. A Survey on Ontology Learning Research. Journal of Software, 2006, 17(9), 1837−1847.

[2] Deng ZH, Tang SW, Zhang M, Yang DQ, Chen J. Overview of ontology. Acta Scientiarum Naturalium Universitatis Pekinensis, 2002, 38(5) 730−738.

[3] Michele Banko, Oren Etzioni. The Tradeoffs Between Open and Traditional Relation Extraction. In Proceedings of ACL-08: HLT, 28–36

[4] Eugene Agichtein, Luis Gravano. Extracting Relations from Large Plain-Text Collections. In Proc. ACM. 2000.

[5] M.A. Hearst. Automatic Acquisition of hyponyms from large text corpora. In the proceedings of the 14th international Conference of Computational Linguistics, 1992, 539-545.

[6] P. Cimiano, A. Hotho, Steffen Staab. Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. Journal of Artificial Intelligence Research 24, 2005, 305-309.

[7] P. Cimiano, Steffen Staab. Learning by Googling in ACM SIGKDD Explorations Newsletter, 2004, 6(2), 24-33

[8] A. Culotta and J. Sorensen. Dependency tree kernels for relation extraction. In Proc. ACL. 2004.

[9] Caraballo, S.A. Automatic Acquisition of a Hypernym-Labeled Noun Hierarchy from Text. Brown University Ph.D. Thesis.2004.

[10] R. Wille. Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts. Dordrecht: Reidel, 1982, 445~470

[11] Li Sujian, Lu Qin, Li Wenjie. Experiments of Ontology Construction with Formal Concept Analysis, OntoLex Workshop IJCNLP 2005, 67-75.

[12] Navigli, R., Velardi, P. Learning domain ontologies from document warehouses and dedicated Websites. Computational Linguistics 30, 2004.

[13] Li Baoli, Chen Yuzhong, Yu Shiwen.Research on Information Extraction: A Survey. Computer Engine and application. 2003.