基于网络挖掘的实体关系元组自动获取

李维刚,刘 挺,李 生
(哈尔滨工业大学计算机学院信息检索研究室,黑龙江哈尔滨 150001)

摘 要: 二元实体关系元组可以应用到知识库构建,数据挖掘,模式抽取等多个领域.本文利用特定关系的一个元组和一个关键词作为种子,结合多种自然语言处理底层技术,采取改进的模式获取方法和自举迭代策略,提出了一种新的从 Web 上抽取实体关系元组的方法.基准方法的平均准确率达到了 78.12%,采用过滤措施后抽取方法的平均准确率达到了 98.42%.实验结果表明,利用网络挖掘方法获取的实体关系元组能够很好满足信息抽取的应用,对抽取出的元组进一步处理,能够获取更多有价值的信息.

关键词: 自举方法; 实体关系; 元组; 信息抽取; 网络挖掘
中图分类号: TP391.2 **文献标识码:** A **文章编号:** 0372-2112 (2007) 11-2111-06

Automated Entity Relation Tuple Extraction Using Web Mining

LI Wei-gang, LIU Ting, LI Sheng
(*Information Retrieval Laboratory, School of Computer Science and Technology, Harbin Institute of Technology, Heilongjiang, Harbin* 150001, *China*)

Abstract: Binary entity relationship tuples can be applied in many fields such as knowledge base construction, data mining and pattern extraction and so on. A seed with a tuple and a keyword of a special relation is used to implement the method of extracting entity relation tuples from the web. Multiple Natural Language Processing (NLP) technologies are combined in this method. A novel pattern acquisition method and an improved bootstrapping iteration strategy are adopted to extract tuples. The baseline method achieves to 78.12% of average precision. The method with filtering measure achieves to 98.42%. The experimental results show that it can satisfy information extraction application well and the extracted tuples can derive more valuable information through further processing.

Key words: bootstrapping; entity relation; tuples; information extraction; web mining

1 引言

信息抽取就是将无结构化的信息转换为结构化或者半结构化信息的过程.目前大部分信息抽取系统是从文本中抽取特定的实体信息,包括时间、机构、地点等.互联网上不仅蕴含了大量的实体信息,还蕴含着实体之间关系的信息.这些实体关系信息能够帮助人们更方便的获取知识,在自动问答、本体构建等领域起着重要的作用,有关实体关系抽取的研究受到越来越多研究者的关注[1~5].

近几年美国国家标准技术研究院(NIST)组织了多次自动内容抽取(Automatic Content Extraction, ACE)评测,其中一项重要评测内容就是实体关系抽取的评测.根据 ACE 所公布的数据集合以及评测标准来看,一般情况下研究者是将实体关系抽取看作一种分类问题,即通过实体对的特征来判断该实体对属于某一类关系类型[6,7],这类方法需要人工标注大量数据.还有一部分学者通过给定关系的若干种子,在有限规模的文本中抽取实体之间的关系.比如,抽取组织和其总部所在地之间的关系、作者和其作品之间的关系等[8~11].这些方法都需要获取若干个初始种子,这需要相对较多的人工劳动,而选择有效的种子也比较困难,并且在迭代过程中容易产生循环依赖问题.所谓循环依赖就是"不好的"实体可能生成无关的模式,而无关的模式在下一次迭代中会生成更多错误元组,这种现象称之为循环依赖.Sekine 在有限文本上实现了一个信息抽取系统 ODIE,该系统利用了模式发现、复述发现等多项技术[2].

上述大部分工作都是在有限文本上进行的,还有一些工作直接利用 Web 作为原始文本资源.Downey 等人从互联网上抽取一元文本模式[12],Szpektor 等人利用一

收稿日期:2006-11-23;修回日期:2007-07-09
基金项目:国家自然科学基金(No.60503072,No.60575042)

个动词词表,针对每一个动词直接从网络来获取相关的句法蕴含关系[13].Ravichandran 等人根据问答系统的多个问题和答案对,从 Web 上抽取答案模板[10].最为著名的一个信息抽取系统 KnowItAll[14],目标是抽取一元关系的实体.Rosenfeld 等人在 KnowItAll 的基础上构建了一个无指导 Web 信息抽取系统 URES[3],其输入和 KnowItAll 类似,需要输入待抽取关系类型、若干个种子实例等.Feldman 利用命名实体识别技术和分类技术增强了 URES 的性能[1].

基于以上分析,本文利用一个简洁的关系定义描述,改进了传统的模式抽取方法和自举迭代策略,结合多种自然语言处理底层技术,提出了基于网络挖掘的实体关系元组自动获取方法.

2 方法概述

本文方法基于这样一个观察:搜索引擎随着查询输入的稍微不同而导致搜索结果具有较大的差异.这是由于互联网上的信息极大丰富,随着查询的不同,很容易找到匹配度更高的结果.另外,和相同的关系关键词以及相同的上下文共现的命名实体(Named Entity,NE)对倾向于具有相同的关系.本文针对搜索引擎的这个特点,利用一个种子元组和关系关键词,通过裁剪方法获取匹配种子的多个上下文模式,然后利用模式再到 Web 上不断的迭代获取更多的相关元组.抽取过程如图 1 所示.

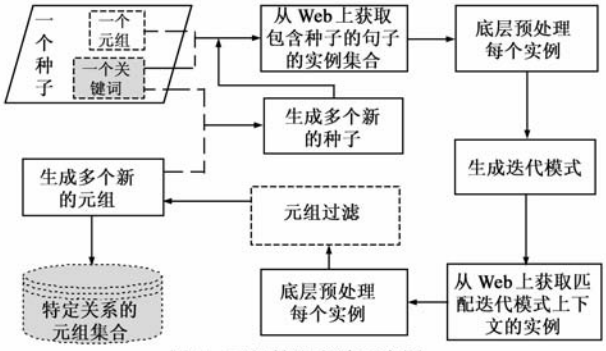

图 1 元组抽取方法示意图

抽取方法的输入是一个简易的种子,输出是一个元组列表.图 1 中虚线框内的"元组过滤"模块表示在本方法中是可配置的,不同的配置可以获取满足不同应用的结果.

3 方法详述

3.1 种子选择

一般的,关系被定义为两个实体或者部分之间的抽象或者从属特征.本文通过一个实例中的两个 NE 和一个关键词来描述该实例的关系类型.两个 NE 构成一个元组,NE 可以是不同类型,也可以是相同类型.一个关键词和两个 NE 一起要能够确定特定的关系类型.满足上述要求的任何种子都可以作为本文方法的输入.表 1 给出了四种关系的种子.

表 1 关系种子

关系类型	种子		
	NE1	NE2	关键词
校长关系	王树国#Nh	哈尔滨工业大学#Ni	校长
首都关系	中国#Ns	北京#Ns	首都
总统关系	布什#Nh	美国#Ns	总统
市长关系	哈尔滨市#Ns	石忠信#Nh	市长

每一个 NE 附带的"#"后的字符串表示 NE 类型信息,其中"Nh,Ns,Ni"分别表示人名、地名和机构名.每一种关系的种子不是唯一的,只需选择满足关系类型的不同 NE 对和关键词就可以构造出不同的种子.可见本文种子选择方法具有构造简单、容易获取的特点,可以方便的从一种关系转换到另一种关系,具有良好的扩展性.

3.2 获取迭代模式

获取迭代模式是元组抽取的重要步骤,其过程是完全自动的,不涉及人工干预,因此具有节省人工劳动的优点.所谓迭代模式就是利用种子从互联网获取的模式,这些模式能够用来获取新的元组.本文提出一种新的模式获取方法,要求迭代模式满足下列条件:

(1)必须包含两个 NE 类型和一个关键词;

(2)至少含有一个实词作为模式的上下文;

(3) 单词总数不能超过一定阈值.

条件(1)保证了抽取出的句子含有正确的关系类型;条件(2)保证了利用搜索引擎能够得到有区别力的结果.由于搜索引擎对于大多数助词、虚词都做了省略处理,若模式上下文中没有实词则不能抽取出有区别力的结果.本文中规定实词类型包含动词、名词和形容词;条件(3)则是在模式上下文的特殊性和抽象性之间进行一个平衡.比如"X#Ni 校长/n#O Y#Nh 认为/v#O"、"X#Ni 校长/n#O Y#Nh 教授/v#O"就是校长关系的两个迭代模式.

假设给定的种子表示为(NE1,NE2,Kw),其中 NE1 和 NE2 分别表示两个 NE,Kw 表示关系的关键词,则具体的获取方法如下:

(1)将种子中的三个元素组合成 NE1 + NE2 + Kw(不包括 NE 类型信息)的形式,作为查询输入到搜索引擎,得到检索结果的标题和摘要(snippet),去除 HTML 标签并将文本分句;

(2)利用字符串匹配技术,只保留那些完全包含种子中三个元素的句子,组成候选句子集合;

(3)对候选句子进行分词、词性标注和 NE 识别处

理,并根据种子中两个 NE 的类型进行过滤,保留满足 NE 类型的句子,最后将从这些句子中获取迭代模式.

本文利用裁剪的方法获取迭代模式,其基本思想就是除了保留种子中的两个 NE 和一个关键词之外,还要获取一个具有区别力的实词作为模式的上下文.一个句子对应一个裁剪后的候选模式.裁剪规则用公式表示如下:

$$P_i=\begin{cases}E[L_{\min}:L_{\max}]+E[L_{\max}:L_{rc}]\\ \quad \text{if } D_{\text{abs}}=1 \text{ and } L_s>L_{kw};\text{or } L_s<L_{kw}<L_b\\ E[L_{lc}:L_{\min}]+E[L_{\min}:L_{\max}]\\ \quad \text{if } D_{\text{abs}}=1 \text{ and } L_b>L_{kw}\\ E[L_{lc}:L_{\min}]+E[L_{\min}:L_{\max}]+E[L_{\max}:L_{rc}]\\ \text{if other conditions}\end{cases}$$

其中,

$$D_{\text{abs}}=\text{abs}(L_{\text{NE1}}-L_{\text{NE2}})$$
$$L_s=\min(L_{\text{NE1}},L_{\text{NE2}})$$
$$L_b=\max(L_{\text{NE1}},L_{\text{NE2}})$$
$$L_{\min}=\min(L_{\text{NE1}},L_{\text{NE2}},L_{kw})$$
$$L_{\max}=\max(L_{\text{NE1}},L_{\text{NE2}},L_{kw})$$

L_x 表示 x 在句子实例 E 中的位置,$x\in\{\text{NE1},\text{NE2},\text{kw}\}$;$D_{\text{abs}}$表示两个 NE 之间的距离;$L_{rc}$和 L_{lc}分别表示模式中右侧和左侧距离 NE 最近的一个实词的位置.$E[m:n]$表示句子实例中从第 m 个词到第 n 个词之间的所有单词,区间取左闭右开;公式中的"+"表示连接的意思,将两个或多个连续片段连接在一起.

需要说明的是,D_{abs}值为 1 的时候,表示两个 NE 在实例中的位置是相邻的.将每个模式中的 NE 全部泛化成变量,同时保留 NE 的类型信息.这样不同的实例裁剪泛化以后可能得到相同的模式.统计模式的出现频率,可以根据获取元组的要求,通过限制不同的频率阈值来保证获取模式的准确性.本文目的是从互联网上挖掘出尽可能多的元组,因此对于迭代模式不做过滤,全部进入迭代过程.

3.3　获取元组

假设校长关系的一个迭代模式为"X#Ni 校长/n#OY#Nh 认为/v#O",具体元组抽取方法如下:

(1)首先获取模式中的上下文,并将其组合在一起作为查询输入到搜索引擎.上面例子所构造的查询为"校长+认为";

(2)从搜索引擎中获取包含所有上下文片段的句子.其中一个句子为:"美国耶鲁大学校长理查德·莱文认为—教学方法影响创新能力培养";

(3)对保留下的句子进行分词,词性标注和 NE 识别处理.然后利用模式匹配分析后的句子,能够匹配的 NE 必须在位置、NE 类型以及上下文都要和模式严格匹配.

美国耶鲁大学/ni#Ni 校长/n#O 理查德·莱文/nh#Nh 认为/v#O -/m#Nm 教学/n#O 方法/n#O 影响/v#O 创新/v#O 能力/n#O 培养/v#O

上面的例子最终抽取出元组(美国耶鲁大学,理查德·莱文).依此方法,可以从实例集合中抽取出多个元组.

3.4　元组的可信度

由于网络信息的复杂性,不可避免的会抽取出噪声模式和元组.减少噪声的方法可以从两个方面考虑,选择可信的模式或元组.由于本文方法的一个特点是只需较少的人工干预,而对迭代模式可信度的计算需要更多的人工知识,这一定程度的抵消了本文方法的这个优点.为此,本文对于迭代模式不进行直接评价,而是对元组的可信度进行了评价.元组可信度自动评价方法描述如下:

(1)首先利用初始种子从 Web 上抽取多个迭代模式,用种子元组实例化每一个模式,然后把每个实例化后的模式作为查询输入到搜索引擎,获取结果中完全匹配查询的估计次数.上节中的例子迭代模式实例化结果为"哈尔滨工业大学校长王树国认为";

(2)将迭代模式按照搜索引擎返回结果的估计次数从大到小排序,取前 n 个模式作为评价抽取元组的模式集合,本文称之为 EvalPSet.求进入 EvalPSet 的每个模式最少出现频率要高于一定阈值,以保证模式具有足够的代表性;

(3)根据 EvalPSet 计算每个元组的可信度,若超过一定阈值,则自动判断该元组为正确的元组.本文实验部分将对多种元组过滤情况进行了评价.可信度的计算公式如下所示:

$$C(T)=\frac{\log\left|\sum_{i=1}^{n}Hits(T+p_i)+\lambda\right|}{\log\left|\sum_{i=1}^{n}Hits(T)+\gamma\right|}\times\sum_{i=1}^{n}\delta_{p_i}$$

其中,

$$\delta_{p_i}=\begin{cases}1, & \text{if } Hits(T+p_i)>0\\ 0, & \text{if } Hits(T+p_i)=0\end{cases}$$

T 是待评测元组,p_i 是 EvalPSet 中一个模式,$T+p_i$ 是指利用元组 T 将模式 p_i 实例化后的结果;λ 和 γ 分别是两个常量,防止计算零的对数.求对数是对搜索引擎返回结果数量的一个平滑,因为实例化后的模式和元组分别到搜索引擎上进行精确匹配的数量往往相差很大,导致两者的点互信息非常小[14].$Hits(x)$是指 x 在搜索引擎上精确匹配结果的估计个数.可信度公式反映了一个元组和集合 EvalPSet 中的模式之间平滑后的点互信息,该值越大,表明元组之间含有正确关系的可

能性越大.

进一步分析,本文元组可信度的评价是通过对唯一确定的初始种子获取的模式进行排序,认为在互联网上出现次数越多则模式越具代表性,构造了模式集合 EvalPSet.任何新抽取出的元组如果和 EvalPSet 中的一个或多个模式匹配,则具有一定的可信度.而错误元组往往是由出现频率较低的模式得到的,通过这种方法可以有效提高元组抽取的准确率.比如元组(许智宏,北京大学)和(钟秉林,北京师范大学)分别为 2.431 和 2.162.

3.5 迭代策略

本文针对互联网的特点,对迭代策略进行了优化,使之能够更好的适应于元组抽取的目的.主要包括以下几点:

(1)为了保证迭代模式的有效性,在获取迭代模式时,除了要求种子中的三个元素都出现在候选实例中,抽取出的迭代模式中还必须包含至少一个实词;

(2)为了尽可能避免迭代过程中的循环依赖现象,本文将抽取出的每一个元组都分别和最初种子中的关键词组合,形成三元组,只有三元组才进入循环过程.这样即使抽取出的元组是错误的,由于该元组和种子关键词共现的概率会非常小,而后续模式获取时要求元组中的两个 *NE* 和种子关键词同时出现在候选句子中,因此能够很好的限制循环依赖问题;

(3)为了保证迭代过程的准确性,本文引入了过滤措施,对抽取出的元组进行可信度的评价,并根据每个元组的可信度决定其是否进入迭代.这是因为一般情况下可信度较低的元组,出现在 Web 上的次数较少,即使进入迭代能够获取的有效信息也有限,限制这些元组可以使迭代过程更快的收敛.

4 实验及分析

4.1 实验设置

由于互联网的特点,和传统方法在有限文本上抽取的评价方法不同,本文只能评价元组抽取方法的准确率,而无法评价完全的召回率.为此,本文针对表 1 所列的 4 种关系类型进行评价,包括校长关系、首都关系、总统关系和市长关系.每种关系所用的种子就是表 1 中所列.本文使用的 NE 识别系统,能够标注 7 种类型的 NE,包括人名、地名、机构名、专有名词、时间、日期和数词短语.

4.2 准确率评测

本文的目标是给定一个特定关系的种子,自动从 Web 上获取满足该关系的大量元组.由于不同关系在互联网上的信息丰富程度不同,因此可能导致最终抽取出的元组数量不同,抽取准确率也不同.对于抽取数量多于 100 个的关系类型,随机选择其中的 100 个元组进行人工评测,对于抽取数量少于 100 的关系类型,将抽取出的全部元组人工评测.

本文分别实现了四种不同配置的抽取方法.第一种为基准方法,在迭代模式获取和元组抽取过程中,不采用任何过滤措施(表示为 B);第二种方法为在元组获取时采用了过滤措施,只保留可信度超过一定阈值的元组进入迭代,而低于可信度阈值的元组直接作为最终结果输出(表示为 B' + F);第三种方法和第二种相反,每一次迭代抽取的元组全部进入迭代过程,最后对获取的所有元组进行过滤,输出可信度高于一定阈值的元组(表示为 B + F');第四种方法结合了第二、三种方法,既过滤迭代元组,也过滤最后的结果元组(表示为 B' + F').

为了综合衡量不同配置方法的性能,本文计算了每种方法的平均准确率,方法如下:

$$P_{\text{average}} = \frac{\sum_{i=1}^{n}(p_i \times N_i)}{\sum_{i=1}^{n}(N_i)}$$

其中,p_i 为第 i 种关系的准确率,N_i 为对应方法抽取出的元组数量.抽取结果如表 2 所示.

表 2 元组抽取结果

关系类型	方法	迭代次数	元组数量	准确率
校长关系	B	9	362	90%
	B' + F	6	311	93%
	B + F'	9	178	100%
	B' + F'	6	169	100%
首都关系	B	6	174	88%
	B' + F	6	170	89%
	B + F'	6	143	95%
	B' + F'	6	142	93%
总统关系	B	9	576	81%
	B' + F	7	432	85%
	B + F'	9	257	98%
	B' + F'	7	249	97%
市长关系	B	12	1079	71%
	B' + F	6	886	84%
	B + F'	12	541	99%
	B' + F'	6	508	100%

从表 2 可以看出,基准方法的最低准确率达到了 71%,最高准确率达到了 90%,而平均准确率达到了 78.12%.这是因为尽管基准方法没有采取任何过滤措施,但是在每次迭代中通过引入关系关键词和每个元组组合在一起进行迭代,该关键词有效的限制了传统自举方法中容易产生的循环依赖问题.

其余三种配置方法引入了过滤措施,获取元组的平均准确率分别为 86.27%、98.42%和 98.37%,与基准

方法相比都有了较大的提高.但在准确率提高的同时,也牺牲了部分召回率,因此最终的抽取数量也有着不同程度的下降.相比较而言,B + F'方法在获取足够高准确率的同时,元组的抽取数量最多,在召回率和准确率上达到了一个较好的平衡.这是因为尽管进入迭代过程的元组有部分错误的元组,这些错误的元组不容易和关键词同时出现,从而限制了错误蔓延.并且,在每次抽取的最后对元组进行基于统计的过滤,在保证抽取元组的数量的同时,可以有效的过滤掉那些错误的元组.

4.3　迭代次数的影响

为了更好的分析迭代次数对于元组抽取准确率的影响,本文对基准方法抽取校长关系的全部结果进行了人工评测.为了直观的表达元组抽取结果,将抽取结果用图形方式显示了出来,如图 2 所示.

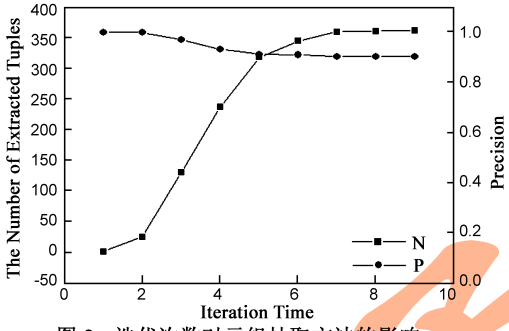

图 2　迭代次数对元组抽取方法的影响

其中曲线 N 为抽取出的元组总体数量随着迭代次数的变化,曲线 P 表示抽取出的元组准确率随着迭代次数的变化趋势.从图 2 可以看出,元组抽取数量随着迭代次数增加而增加,但是增长趋势在达到一定迭代次数以后逐渐减缓.这是因为在后面的迭代过程中有一部分元组已经被抽取出来,而每次迭代获取的元组数量只计算新抽取出的元组,直至抽取不到新的元组,迭代过程自动结束.

从 P 曲线上可以看出,准确率下降趋势明显的地方,是在某一次迭代过程中获取的个别错误迭代模式,由错误迭代模式抽取出的错误元组,从而导致抽取出的元组准确率下降较快;但是准确率的下降趋势并没有随着迭代次数的增加而一直增加,这是因为抽取出的元组需要和相应关系的关键词结合以后才作为下一轮的迭代种子,这很大程度上限制了错误蔓延的情况.

4.4　错误分析

为了更清楚的阐明抽取方法的特点,本文还进行了错误分析.首先将每种方法在四种关系中出现的所有错误元组集中在一起,全部人工判断其错误类型,结果如表 3 所示.表中的数据分别为特定错误类型的个数和其在所有错误中所占的比例.I 型为 NE 识别错误引起的错误,II 型为迭代模式识别错误引起的错误,III 型为互联网噪声信息引起的错误.

表 3　错误分析结果

方法	错误个数(所占比例)		
	I 型	II 型	III 型
B	25(35.7%)	43(61.4%)	2(2.90%)
B' + F	23(46.9%)	25(51.0%)	1(2.10%)
B + F'	3(37.5%)	4(50.0%)	1(12.5%)
B' + F'	3(30.0%)	6(60.0%)	1(10.0%)

一般情况下,I 型错误都是含有部分正确的元组,比如首都关系的一个元组(古巴,哈瓦那老城),其对应的正确元组分别为(古巴,哈瓦那);II 型是由迭代模式识别错误引起的,如校长关系的一个元组(靳润成,郑州大学)就是一个错误的元组,它是根据模式"校长 Nh 一行到 Ni 参观"抽取出来的,而这个模式是从实例"天津师范大学校长靳润成一行到郑州大学参观访问"中得到的;III 型错误是由于部分网络信息的非权威性造成的,这一类信息往往包含着错误的信息,比如首都关系的一个元组(丹麦,伦敦)就是从实例"让伦敦成为丹麦的首都吧!"中获取得到的,显然这个元组是错误的.

从表 3 可以看出,在基准方法和"B' + F"方法中,I 型和 II 型错误占了绝大部分;在后两种方法中,通过对最终抽取结果进行过滤,三种类型的错误都得到了有效的遏制.在未来工作中可以通过 NE 识别模块的性能提升、引入更多的知识对迭代模式进行过滤等方法获取更为准确的元组.

4.5　召回率的分析

上文主要对元组抽取方法的准确率进行了评测,对于召回率则没有进行详细的评测,这是由于互联网上的信息特点决定的.本文通过对影响元组抽取数量的分析,间接的对召回率进行了描述.通过分析得出,影响最终元组抽取数量的主要原因除了迭代模式的有限性之外,还有以下两个因素:

(1)NE 识别的不准确性;从本文的介绍来看,NE 识别技术在本文的工作中起着举足轻重的作用,如果一个 NE 识别错误,则和该 NE 相关的元组一定抽取不出来.因此 NE 识别的准确率对于本文方法最终抽取出的元组的数量和质量都起着决定性的作用;

(2) NE 之间的指代问题;一个地名 NE 可能有多种形式.比如抽取首都关系时,地名 NE"沙特阿拉伯",还可能表示为"沙特"、"沙特阿拉伯王国"等多种形式.指代问题对于元组的匹配数量也起到一定的影响.

5　结论

本文利用一个元组和关键词作为种子,结合多种自然语言处理底层技术,从 Web 上自动抽取特定关系

的大量元组.和已有方法相比主要有以下贡献:

(1)种子选择方法:种子是包含一个元组和一个能够确定两个 NE 之间关系类型的关键词,容易构造,只需较少的人工劳动,具有良好的可扩展性;

(2)迭代模式获取方法:针对互联网上信息极大丰富的特点,本文改进了模式获取的方法,不关注迭代模式的完整性,而是引入具有区别意义的上下文进入模式;

(3)改进的自举迭代策略:通过统计互联网上的信息对获取的元组进行可信度评价,并将新获取的元组加入"关系关键词"的限制,有效降低了发生循环依赖的可能性.

实验结果显示,本文基准方法抽取元组的平均准确率达到了 78.12%,引入过滤措施之后,获取元组的平均准确率达到了 98.42%,能够很好的满足信息抽取的应用.

本文利用网络挖掘方法对元组抽取进行了初步探索,拟在以下几个方面进行更为深入的研究:对迭代模式进行可信度评测,从而获取更为准确的元组;引入句法和词性信息,不单独的依赖 NE 识别结果,扩展到抽取一般名词和名词短语上去;利用抽取出的元组进行更进一步的应用研究,包括模式的获取、Ontology 的构建等.

参考文献:

[1] R Feldman, B Rosenfeld. Boosting unsupervised relation extraction by using NER[A]. Conference on Empirical Methods in Natural Language Processing[C]. Australia: BPA Digital, 2006.473 – 481.

[2] S Sekine. On-demand information extraction[A]. COLING/ACL 2006 Main Conference Poster Sessions[C]. Australia: BPA Digital, 2006.731 – 738.

[3] B Rosenfeld, R Feldman. URES: an unsupervised web relation extraction system[A]. COLING/ACL 2006 Main Conference Poster Sessions[C]. Australia: BPA Digital, 2006.667 – 674.

[4] M Zhang, J Su, et al. Discovering relations between named entities from a large raw corpus using tree similarity-based clustering[A]. The 2nd International Joint Conference on Natural Language Processing[C]. Berlin Heidelberg: Springer-Verlag, 2005.378 – 389.

[5] T Hasegawa, S Sekine, R Grishman. Discovering relations among named entities from large corpora[A]. Meeting of the Association for Computational Linguistics[C]. New York: ACM Press, 2004.415 – 422.

[6] 车万翔,刘挺,李生.实体关系自动抽取[J].中文信息学报,2005,19(2):1 – 6.
W X Che, T Liu, S Li. Automatic entity relation extraction[J]. Journal of Chinese Information Processing. 2005, 19(2): 1 – 6. (in Chinese)

[7] C Aone, M Ramos-Santacruz. Rees: A large-scale relation and event extraction system[A]. Applied Natural Language Processing Conference[C]. San Francisco: Morgan Kaufmann, 2000.76 – 83.

[8] S Brin. Extracting patterns and relations from the world wide web[A]. WebDB Workshop at the 6th International Conference on Extending Database Technology[C]. London: Springer-Verlag, 1998.172 – 183.

[9] D Ravichandran and E Hovy. Learning surface text patterns for a question answering system[A]. The 40th Association for Computational Linguistics[C]. Morristown: Association for Computational Linguistics, 2002.41 – 47.

[10] R Jones, A McCallum, et al. Bootstrapping for text learning tasks[A]. IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications[C]. San Fransisco: Morgan Kaufmann, 1999.52 – 63.

[11] A Eugene, G Luis. Snowball: Extracting relations from large plain-text collections[A]. ACM International Conference on Digital Libraries[C]. New York: ACM Press, 2000.85 – 94.

[12] D Downey, O Etzioni. Learning text patterns for web information extraction and assessment[A]. AAAI-04 Workshop on Adaptive Text Extraction and Mining[C]. California: AAAI Press, 2004.50 – 55.

[13] I Szpektor, H Tanev, et al. Scaling web-based acquisition of entailment relations[A]. Conference on Empirical Methods in Natural Language Processing[C]. New York: ACM Press, 2004.41 – 48.

[14] O Etzioni, M Cafarella, et al. Unsupervised named-entity extraction from the web: An experimental study[A]. Artificial Intelligence[C]. Netherlands: Elsevier Science BV, 2005.91 – 134.

作者简介:

李维刚 男,1979 年生于河南永城,哈尔滨工业大学计算机科学与技术系博士研究生.主要研究方向为复述技术、信息抽取、自动问答和机器翻译.E-mail: lee@ir.hit.edu.cn

刘 挺 男,1972 年生于哈尔滨,哈尔滨工业大学计算机科学与技术系教授、博士生导师.主要研究方向包括信息检索、自然语言处理和机器翻译.