基于特征向量与 SVO 扩展的企业生态关系抽取

代江波,毛建华,刘学锋,张鸿洋
(上海大学 通信与信息工程学院,上海 200444)

摘 要:企业间关系是企业价值链中最重要的组成部分之一,也是企业管理者最感兴趣的信息之一,对于企业决策和管理具有重大意义;从企业年报文本中抽取企业关系和实体关系要求高时效性和强鲁棒性。实体关系抽取的核心问题在于关系模式的选择和提取,由于中文句式较复杂、表达方式灵活、语义多样等固有性质的限制,导致关系实例的关系表述不准确,语义信息表示不足。因此,提出基于特征向量与 SVO 扩展的企业关系抽取模型,并且在该方法中引入触发词机制,然后使用具有触发词约束的关系模式对年报文本进行企业关系的抽取。最后通过对 1 000 家上市企业的年报文本进行实验,实验结果表明,该方法能较大地提高实体关系的抽取性能。

关键词:企业关系抽取;触发词;特征向量;SVO 扩展;关系模式
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2018)10-0139-06
doi:10.3969/j.issn.1673-629X.2018.10.029

Enterprise Ecological Relationship Extraction Based on Feature Vector and SVO Extension

DAI Jiang-bo, MAO Jian-hua, LIU Xue-feng, ZHANG Hong-yang
(School of Communication & Information Engineering, Shanghai University, Shanghai 200444, China)

Abstract: The relationship between enterprises is one of the most important components in the enterprise value chain, and also one of the most interesting information for business managers, which is of great significance to the decision-making and management of enterprises. It needs high timeliness and strong robustness to extract the relationships from the annual reports. The core problem of the entity relation extraction is the restriction of the inherent nature of the selection and extraction of the relational model. Due to the complex Chinese sentence, the flexible expression and the variety of semantics, the representation of relational examples is inaccurate and the semantic information is insufficient. Therefore, we propose a model of enterprise relationship extraction based on feature vector and SVO extension. In this method, the triggering mechanism is introduced, and then the relational model with the triggering words constraint is used to extract the relationship of the enterprise from the annual report text. Finally, the annual report text of 1 000 listed companies is tested. The experiment shows that this method can greatly improve the extraction performance of the entity relationship.

Key words: enterprise relationship extraction; triggering words; feature vector; SVO extension; relational model

0 引 言

企业实体关系抽取旨在从自然语言文本中抽取企业实体及企业实体间关系,以帮助企业获取企业自身或其他感兴趣企业的相关信息。网络文本,尤其是上市公司年报定期披露的上市公司生产经营的相关信息中包含丰富的企业产品生产、采购、销售、市场竞争、行业发展及政策法规等内容,但上市公司定期报告披露的企业关系信息是破碎的。为构建行业企业的关系图谱,有必要研究行业企业关系的抽取方法。企业实体关系的获取是一种典型的信息抽取问题,主要是研究在实体识别的基础上确定文本中实体对所蕴含的关系类型。

然而对于实体关系抽取,企业实体间关系的抽取有如下特殊性:缺乏大规模企业实体关系标注语料;存在企业实体间关系模式的表达与选取问题。由于仅以词法、句法特征抽取实体关系不能获取复杂句子的长距离深层特征,而最短依存路径模式抽取实体关系由于实现了对关系的浓缩表示缺乏语义约束以及相关特

收稿日期:2017-11-21 **修回日期**:2018-03-15 **网络出版时间**:2018-05-28
基金项目:国家自然科学基金(61271061);上海市自然科学基金(16ZR1411100)
作者简介:代江波(1991-),男,硕士研究生,研究方向为文本事件抽取与分析;毛建华,博士后,副教授,研究方向为文本事件抽取与分析;刘学锋,博士后,教授,研究方向为遥感与空间信息处理。

征表示容易造成关系实例误检。基于以上问题,文中重在解决如何在缺乏标注语料的前提下,实现企业间实体关系模式的建立及其关系的抽取。

1 相关工作

近年来的研究趋势表明,机器学习方法是目前关系抽取研究的主流方法,主要包括有监督、无监督和半监督三种类型。

有监督方法是通过构建一个监督分类器来实现关系抽取。其主要思想是通过对人工标记的训练样本训练出关系实例的分类模型,然后利用训练出的分类模型测试候选关系实例集合实现对实例关系的分类。在有监督的方法中,基于树核的方法不需要构造特征向量,解决了基于特征向量方法无法充分利用实体对上下文的结构信息的问题。文献[1]将改进的语义序列核函数和机器学习 KNN 分类算法相结合构造分类器,并对关系类型进行分类和标注;在 ACE 数据集上,关系抽取的平均准确率提高到 88%。文献[2]首先构造一个丰富的语义关系树结构,将句法信息和语义信息相结合,并对上下文相关的树结构进行扩展。实验表明提出的基于树核的方法优于其他先进的方法。文献[3]在实例句的句法树中融入能反映特定领域实体语义关系的领域知识树,相比没有融入领域信息的方法实体关系抽取的性能提高了 3.4%。文献[4]通过抽取深层句法特征构建演化关系模式并借助 CRF 模型识别概念间的演化关系,使演化关系的抽取较传统方法有更高的准确性。该模式具有一定的优越性和有效性,可以有效识别机器学习领域中概念间的演化关系。然而有监督方法训练和测试速度过于缓慢,不适合处理大规模数据。

无监督实体关系抽取方法实际上是一种聚类算法,将相似度高的实体对所在的关系句子实例聚为一类,然后选择评分高并且具有代表性的特征词来标记这种关系。无监督方法无需依赖疏通关系标注的语料,因此可以解决有监督方法中标注困难的问题。文献[5]提出了一种能同时触发关系触发词和关系参数的无监督的生成模型,并通过抽取的实体关系揭示了患者记录中存在的一些隐含的语义结构。文献[6]提出一种基于多层无监督神经网络的分类模型,有效解决了高维特征向量中特征提取以及分类的问题。实验表明在高维空间特征的信息抽取任务中,DBN 模型具有较强的处理和分析能力,其效果比 SVM 和反向传播网络更好。文献[7]提出了一个多级聚类方法来分组语义等价关系,该方法不仅提高了可扩展性,而且通过利用每个初始簇中的冗余来提高聚类结果。无监督实体关系抽取方法在处理大规模实体关系抽取时虽然有一定的优势,但其聚类阈值难以提前确定,并且目前仍缺乏较客观的评价标准。

由于有监督的关系抽取方法需要大量的标注语料库,而无监督的机器学习其聚类阈值难以事先确定,并且目前仍缺乏较客观的评价标准。所以为了减少对语料集的人工标注,实现关系的自动抽取,半监督的关系抽取算法得到了发展和应用。2014 年,文献[8]提出一种改进的上下文模式语义分析并结合基于 bootstrapping 的半监督算法抽取语义关系抽取,在一定程度上加强了语义关系抽取的效果。文献[9]提出一种基于词嵌入的 bootstrapping 关系抽取模型,并且依靠词嵌入实现了从一组新闻线文档中提取四种关系的任务,获得了较好的表现。文献[10]定义了一种语义约束的 bootstrapping 关系抽取模型,并提出了语义最短依存路径关系模式语义,使其包含了更丰富的句法特征和语义特征,具有更强的关系指向性,且最终具有较好的表达效果。

基于以上研究,文中建立 bootstrapping 关系抽取模型以减少对语料集的人工标注,扩充种子关系模式集合,实现关系模式的自动抽取;在关系抽取模型中提出基于触发词的特征向量(T_FVM)关系模式和基于触发词的主谓宾扩展(T_SVOE)关系模式,以解决企业关系抽取中对表格信息处理和对句子语义信息表示不足的问题。

2 基于特征向量与 SVO 扩展的企业关系抽取

企业关系抽取主要包含预处理、构建种子集、迭代和测试四个子模块。预处理模块是对企业年报原始文本进行正文抽取、企业专有名词识别以及触发词构建。构建种子集则从语料集中选取具有代表性的一部分关系实例进行标注。迭代过程是 bootstrapping 的核心和重点,首先对训练语料集进行依存句法分析、特征提取[11-13]等产生候选关系模式,然后对候选模板进行相似性分析与评价,将可靠的关系模式保留下来扩展种子集合,最后将扩展的种子集合作为下一次迭代的输入。将得到的关系模式集合对测试预料中的实例进行关系抽取。其核心流程如图 1 所示。

在候选关系模式模块中主要运用提出的基于特征向量与 SVO 扩展的关系模式,并在关系模式中引入触发词语义约束机制。

2.1 构建企业关系专有名词

企业关系专有名词是指能够触发一定关系类型的词,具有一定的语义指向性,也被称为关系指示词或触发词。关系指示词常在关系抽取中被用作实体关系发生的指向词;在关系抽取中,主要是指具有某种语义关

系并能触发特定关系模式的词[14-15]。

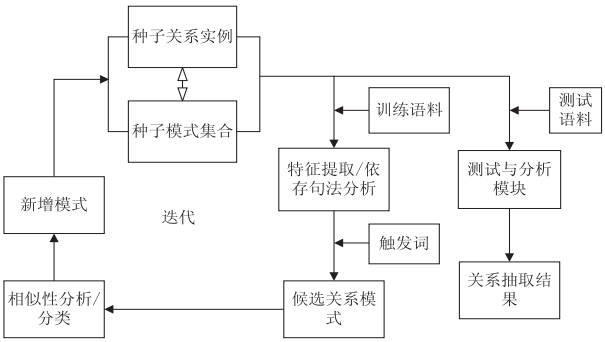

图 1　企业实体关系抽取核心流程

通过对关系模式添加触发词特征以实现语义关系的约束和实体关系的准确表达。它作为关系模式的语义锚点,直接关系着关系模式的语义类型,对关系抽取起着重要的作用。文中利用统计方法和人工筛选、添加的方法实现触发词的获取和过滤。定义了 5 种实体关系:客户关系、供应商关系、研发关系、附属关系、位置关系;各关系的具体定义为:

定义 1(客户关系):指企业为达到其经营目标,主动与客户建立起销售与购买的联系。

定义 2(供应商关系):指企业为达到其经营目标,主动与供应商建立起购买与销售的联系。

定义 3(研发关系):指企业为达到其经营目标,主动提高技术、产品和服务水平,将研究成果转化为可靠,具有成本效益的创新产品的活动。

定义 4(附属关系):存在一定隶属关系或合作关系的两个或两个以上组织、机构或者企业。

定义 5(位置关系):是指组织、机构以及公司等与地名或地址等存在的特定关系。

针对描述每一种企业关系的专有名词的具体定义情况如表 1 所示。

表 1　企业关系专有名词定义

关系类型	企业关系专有名词列表
客户关系	客户 顾客 销售 出售 发售 提供 供给 供应
供应商关系	供应商 采购 买 收买 购买 收购 购入 购得 购价 买购 竟购
研发关系	在研 研发 研制开发 研究 研制
附属关系	全资子公司 授权 控股子公司 参股 投资 注资 控制 下属 子公司 拥有 旗下子公司 下属子公司 经营 托管 掌管 控制 所属 附属　隶属 属 直属 隶属于 属于 收购　并购　整合 收购 投入 合并
位置关系	坐落 位于 地点 注册地点 注册地 地方 坐落于

2.2　关系模式及其表述

关系模式的表达和抽取是实体关系抽取的核心问题,其目前主要是对关系的向量表示和结构化表示。由于关系选择性与关系模式息息相关,好的关系模式具有好的关系选择性,可以提高关系抽取的正确率,因此关系模式的表达和抽取成为实体关系抽取的关键。文中提出基于触发词的特征向量关系模式(FVM based on semantic constraint of trigger words,T_FVM)和触发词的主谓宾(SVO)扩展关系模式(SVO extension model based on semantic constraints of trigger words,T_SVOE)进行实体关系抽取。

2.2.1　基于触发词的特征向量关系模式

对于企业语料中的半结构化表格的处理,需要采取间接方法来获取相应的实体关系;这一部分主要是抽取年报公司实体和表格中企业、组织、机构以及产品实体间的关系而言,对于表格内部实体间的关系并不能有效表示。

表格中的企业实体由于缺乏信息及特征以至于实体间的关系难以表达和获取;但是在企业年报中凡是和年报公司实体含有一定实体关系的表格都会提前对相关表做一个关键性的相关信息描述,因此可以对这部分信息进行详细的分析和信息特征获取。由于这一部分依存句法特征较少,分析结果不理想,因此这一部分采用基于特征向量的关系模式(FVM)进行表示,并使用触发词进行语义约束,即提出基于触发词语义约束的特征向量关系模式(T_FVM)。

定义 6:定义四元组<ER;ES1;ES2;FV>为一个完整关系实例。其中,ER(entity relation)为两个实体间存在的关系,例如客户关系、供应商关系、研发关系等;ES1、ES2(entities)即为含有一定关系的实体对;FV(feature vector)为特征集合即关系实例的关系表述。种子实例中对于特征向量关系模式抽取的关系实例也以这样的四元组形式表现。

文中在传统浅层词汇特征的基础上,增添触发词特征和实体类型特征以获取实体对之间更丰富的关系特征,并使用 KNN 算法进行分类预测。选取的实体关系特征有:

(1)关键词汇特征序列。首先使用 TextRank[16] 算法获取关键词汇特征序列,TextRank 算法是利用局部词汇之间的关系(共现窗口)对后续关键词进行排序,直接从文本本身抽取。公式定义如下:

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \tag{1}$$

其中,d 为阻尼系数,取值范围为 0 到 1,表示从图中某一特定点指向其他任意点的概率,一般取值为 0.85。

(2)触发词特征。通过已构建的企业关系专有名词库获取句子中含有的触发词特征。

(3)实体类型特征。能够识别的实体种类有人名、地名、组织机构名等。

2.2.2 基于触发词的主谓宾扩展关系模式

对于纯文本信息,则将含有命名实体的句子作为关系抽取的元实例集,这一部分则采用 T_SVOE 模式进行模式表示以及关系抽取。对于 SVO[17-18]模式,虽然可以利用主谓宾组合表示一定的深层语义信息,但对于长复杂句,由于信息表现过于简单,容易造成对句子关键信息表述的缺失。例如句子“嘉园环保为本公司全资子公司,注册资本 6 000 万元”,如果是基于 SVO 模式,则提取的基本句子关系模式为:“嘉园环保:company_name SBV 注册资本 * 元”;则具有语义表述的关键信息成分“控股子公司”出现缺失,结果导致本句子被归类为无关系。

因此文中提出 SVO 扩展模式;SVO 扩展模式的重点在于对主谓宾各个成分含有的并列关系(COO)、定中关系(ATT)以及介宾关系(POB)的句子成分进行提取,并将其依存句法关系也加入结构模式中使之可以完善对句子语义信息的表达。由于关系模式抽取时可能会产生语义漂移问题,因此使用触发词进行语义约束即提出基于触发词的 T_SVOE 关系模式。

依存语法通过分析语言单元中各成分之间的依存关系,即指出句子中单词之间的句法搭配,分析句子的主谓词,宾语,形式,补语结构,揭示句子成分之间的语义修饰关系。直观地讲,依存句法分析句子中的这些语法成分与这些成分的位置无关,分析各成分之间的语义修饰关系,可以获得远距离的搭配信息。常用的依存句法分析标注关系如表 2 所示。

表 2 依存分析标注关系

关系类型	关系标注	例子
主谓关系	SBV	嘉园环保注册资本 6 000 万元(嘉园环保←注册)
动宾关系	VOB	注册资本 6 000 万元(注册→元)
介宾关系	POB	为全资子公司(为→全资子公司)
并列关系	COO	上海康数和苏南互联网(上海康数→苏南互联网)
核心关系	HED	指整个句子的核心
定中关系	ATT	全资子公司(全资←子公司)

定义 7:定义五元组<ER;ES1;ES2;SVOE;TW>为一个关系实例。其中 ER 为两个实体间存在的关系,例如客户关系、供应商关系、研发关系等;主谓宾扩展(SVOE)模式即对主谓宾各个成分含有的并列关系、定中关系以及介宾关系的句子成分进行提取,并将其依存句法关系也加入结构模式中使之可以完善对句子语义信息的表达;ES1、ES2 即为含有一定关系的实体对;TW 即触发词在句法中对实体关系具有指示作用,在句法结构中要保留该元素,如果句子中含有触发词则 TW 为触发词名称,反之为 No。在种子实例中,针对 T_SVOE 的关系模式抽取的关系实例也是以这样五元组形式表现的。

根据依存句法分析获取 T_SVOE 关系模式主要包括三步:

(1)依据依存句法分析结果,提取包含实体以及触发词在内的主谓宾核心结构,如算法 1 所示。

(2)提取与主谓宾有直接并列关系、定中关系以及介宾关系的依存关联节点,如算法 2 所示。

(3)加入依存关系特征并用实体类型替代实体部分,其他无关的成分用 * 代替即得到最终可以表示一定实体关系的关系模式。

算法 1:提取包含实体以及触发词在内的主谓宾核心结构。

输入:实体 e,实体所在句子的依存句法分析结果

输出:包含实体以及触发词在内的主谓宾核心结构

Foreach node sentenceNodes

IF(node→relate 为 SBV,node→relate 为 HED,node→relate 为 VOB 或者 node→name 为 e)

/ * 遍历依存句法中每个关系节点,提取主谓宾关系成分到节点关系集合中 * /

Add node→relate to nodeSet

ENDIF

Foreach tw triggerWordsSet

/ * 遍历触发词集合,获取句子关系中含有触发词的关系成分 * /

IF(node→name 为 tw)

Add node→relate to nodeSet

ENDIF

END

END

算法 2:提取与主谓宾的直接依存关联节点。

输入:主语 S、谓语 R、宾语 O 成分以及实体所在句子的依存句法分析结果

输出:主语 S、谓语 R、宾语 O 的依存关联节点集合

p=S,R,O←parent;

Foreach p ∈ sentenceNodes

IF(p→relate 为 COO,p→relate 为 ATT 或者 p→relate 为 POB)

Add p tonodeSet

/ * 遍历依存句法中每个关系节点,如果与主语 S、谓语 R、宾语 O 的依存关系为 COO、ATT 或 POB,则添加关系节点到节点集合中 * /

ENDIF

END

则句子“嘉园环保为本公司全资子公司,注册资本 6 000 万元”通过基于 T_SVOE 的关系模式分析得到 T_SVOE 关系模式为“嘉园环保:company_name SBV

为 VOB 全资子公司 * ,注册资本 * ";通过和基于主谓宾(SVO)模式的结果对比分析,基于触发词和 SVOE 模式明显对句子有更完善的语义信息表达。

3　实　验

3.1　实验数据集及预处理

选取 2015 年 1 000 家上市公司年报作为语料,其中涉及电子、汽车、药业、食品、房地产五大行业各 200 篇左右,并定义 5 种实体关系:客户关系、供应商关系、研发关系、附属关系、位置关系。文中采取半监督的方法,首先构造一个小的种子集,然后从未标记的数据集中提取某些特征的数据,在评估之后,提取最可信的数据集来扩展训练集,然后迭代这个过程,因此不需要进行大量标注构建小规模的种子集合是关键。对每个行业年报选取 20% 用来建立种子集,其余 80% 中的 60% 作为训练语料,40% 作为测试语料。

预处理是实体关系识别的基础也是关键工作。选取 FudanNLP 自然语言处理工具进行分词、停用词过滤,词性标注以及实体识别等基础工作,然后获取包含实体的句子,使用 LTP-Cloud 进行句子结构解析以及词之间的依存分析。

3.2　实验结果评价指标

准确率(Precision, P_R)、召回率(Recall, R_R)和 F 值(F -Measure, F)是信息抽取领域中的三项基本评价指标。准确率是指在某一特定关系类型的实例中被正确抽取的实例占所有抽取为此关系类型的比例,它是从查准率的角度评估抽取效果;召回率是指在某一特定关系类型的实例中被正确抽取的实例占实际属于本类型的实例的比例,它是从查全率的角度评估抽取效果;由于准确率和召回率都是单一方面的评估,实际两者是相互影响、相互牵制的,因此 F 值则是综合考虑准确率和召回率的影响。计算公式为:

$$P_R = \frac{\text{被正确抽取的于系 } R \text{ 的实体对的个数}}{\text{所有被抽取为关系 } R \text{ 的实体对的个数}} \tag{2}$$

$$R_R = \frac{\text{被正确抽取的属于于 } R \text{ 的实体对的个数}}{\text{实际应被抽取的属于关系 } R \text{ 的实体对的个体}} \tag{3}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{4}$$

3.3　实验结果与分析

基于触发词的语义模式企业关系抽取任务中,采用基于 T_FVW 和基于 T_SVOE 两种关系模式处理不同情况下的实体关系,同时引用的触发词约束在语义上约束关系模式,以确保新添加的关系模式和关系实例指向当前关系,这大大改善了抽取的准确率。表 3 显示了 5 种实体关系抽取的准确率、召回率和 F 值。

表 3　各关系抽取性能　　%

关系类型	Precision	Recall	F
客户关系	83.2	77.9	80.5
供应商关系	83.6	74.0	78.5
研发关系	80.2	73.1	76.5
附属关系	84.1	71.9	77.5
位置关系	82.3	70.1	75.7

同时还实现引入其他三种关系抽取算法进行对比,使用未加触发词语义约束的 FVW 和 SVO 关系模式,该方法记为 FVM&SVO;使用未加触发词语义约束的 FVW 和 SVOE 关系模式,该方法记为 FVM&SVOE;同时,还实现未扩展 T_SVO 关系模式以方便和扩展 T_SVOE关系模式进行对比,该方法记为 T_FVW&T_SVO;记文中方法为 T_FVW&T_SVOE;对比结果如表 4 所示。

表 4　所有关系下的综合性能比较　　%

方法	Precision	Recall	F
(方法 1)FVM&SVO	67.8	68.2	67.9
(方法 2)FVM&SVOE	71.4	71.0	71.2
(方法 3)T_FVW&T_SVO	79.4	70.7	74.8
文中方法(T_FVW&T_SVOE)	82.7	73.4	77.7

通过比较可以观察到,基于 SVO 扩展的关系模式其准确率和召回率都有小幅度的提升,这说明基于 SVO 扩展模式相比单纯的基于 SVO 模式更适合用于关系的抽取;方法 3 的准确率明显高于方法 2,这主要是因为引入触发词特征对相应的关系模式有一定的语义约束作用,但是召回率下降了 0.3% ,可能是由于关系模式泛化能力不足导致。

观察表 3 和表 4,对于基于关系模式的半监督学习方法具有较高的或快速增长的准确率,相反召回率却增长缓慢甚至偶尔出现较低的现象,这些主要是由 bootstrapping 迭代所产生的,因为该过程是通过候选关系模式集来扩大实例集,然后再通过扩大的实例集反过来扩大关系模式集。如果不能保证抽取的关系模式有较高的准确率,则在迭代过程中必然会导致错误的积累和叠加;因此,基于关系模式的半监督学习方法往往具有较高的准确率。

4　结束语

文中提出了基于特征向量与 SVO 扩展的企业关系抽取模型,并在该模型中引入触发词约束机制。实验结果表明,该方法能够从大规模的企业文本中抽取出企业实体关系,有效解决了企业关系抽取中对表格

信息处理和对句子语义信息表示不足的问题;同时使用 bootstrapping 算法通过种子模板抽取关系模式,不断迭代学习,最终达到需要的数据信息规模,解决了人工干预和语料标注的问题。

下一步将研究跨文档中隐式关系的抽取,以及基于 Web 的企业关系抽取,从而挖掘出更多的实体关系,自动建立全方位的企业生态关系图谱。

参考文献:

[1] 刘克彬,李 芳,刘 磊,等. 基于核函数中文关系自动抽取系统的实现[J]. 计算机研究与发展,2007,44(8):1406-1411.

[2] ZHOU Guodong, QIAN Longhua, FAN Jianxi. Tree kernel-based semantic relation extraction with rich syntactic and semantic information[J]. Information Sciences, 2010, 18(8): 1313-1325.

[3] 陈 鹏,郭剑毅,余正涛,等. 融合领域知识短语树核函数的中文领域实体关系抽取[J]. 南京大学学报:自然科学版,2015,51(1):181-186.

[4] 高俊平,张 晖,赵旭剑,等. 面向维基百科的领域知识演化关系抽取[J]. 计算机学报,2016,39(10):2088-2101.

[5] RINK B, HARABAGIU S. A generative model for unsupervised discovery of relations and argument classes from clinical texts[C]//Proceedings of the conference on empirical methods in natural language processing. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011: 519-528.

[6] 陈 宇,郑德权,赵铁军. 基于 Deep Belief Nets 的中文名实体关系抽取[J]. 软件学报,2012,23(10):2572-2585.

[7] WANG Wei, BESANÇON R, FERRET O, et al. Semantic clustering of relations between named entities[C]//International conference on natural language processing. [s. l.]: Springer International Publishing, 2014: 358-370.

[8] YE Feiyue, SHI Hao, WU Shanpeng. Research on pattern representation method in semi-supervised semantic relation extraction based on bootstrapping[C]//Seventh international symposium on computational intelligence and design. Hangzhou, China: IEEE, 2014: 568-572.

[9] BATISTA D S, MARTINS B, SILVA J M. Semi-supervised bootstrapping of relationship extractors with distributional semantics[C]//Proceedings of the 2015 conference on empirical methods in natural language processing. Lisbon, Portugal: Association for Computational Linguistics, 2015: 499-504.

[10] ZHANG Chunyun, XU Weiran, MA Zhanyu, et al. Construction of semantic bootstrapping models for relation extraction[J]. Knowledge-Based Systems, 2015, 83: 128-137.

[11] 李明耀,杨 静. 基于依存分析的开放式中文实体关系抽取方法[J]. 计算机工程,2016,42(6):201-207.

[12] 郭喜跃,何婷婷,胡小华,等. 基于句法语义特征的中文实体关系抽取[J]. 中文信息学报,2014,28(6):183-189.

[13] 甘丽新,万常选,刘德喜,等. 基于句法语义特征的中文实体关系抽取[J]. 计算机研究与发展,2016,53(2):284-302.

[14] 李培峰,周国栋,朱巧明. 基于语义的中文事件触发词抽取联合模型[J]. 软件学报,2016,27(2):280-294.

[15] 刘绍毓,席耀一,李弼程,等. 无监督实体关系触发词词典自动构建[J]. 计算机应用与软件,2016,33(5):72-76.

[16] LU Guangming, XIA Yule, WANG Jiamei, et al. Research on text classification based on TextRank[C]//International conference on communications, information management and network security. [s. l.]: [s. n.], 2016.

[17] CARLSON A, BETTERIDGE J, WANG R C, et al. Coupled semi-supervised learning for information extraction[C]//Proceedings of the third ACM international conference on web search and data mining. New York, NY, USA: ACM, 2010: 101-110.

[18] MILAJEVS D, SADRZADEH M, ROELLEKE T. IR meets NLP: on the semantic similarity between subject-verb-object phrases[C]//Proceedings of the 2015 international conference on the theory of information retrieval. Northampton, Massachusetts, USA: ACM, 2015: 231-240.

(上接第 138 页)

策略[J]. 计算机工程,2015,41(11):114-119.

[7] 王 来,瞿健宏. 基于 HDFS 的分布式存储策略分析[J]. 智能计算机与应用,2016,6(1):5-8.

[8] 朱刘江. 基于 Hadoop 的海量城市交通流数据分布式存储与分析研究[D]. 扬州:扬州大学,2015.

[9] 洪旭升,林世平. 基于 MapFile 的 HDFS 小文件存储效率问题[J]. 计算机系统应用,2012,21(11):179-182.

[10] 刘晓霞. Hadoop 中大量小文件性能优化方法研究[J]. 计算机光盘软件与应用,2013,16(18):78-80.

[11] 漆 铖. 云环境下海量小文件存储技术的研究与应用[D]. 广州:广东工业大学,2015.

[12] 余 思,桂小林,黄汝维,等. 一种提高云存储中小文件存储效率的方案[J]. 西安交通大学学报,2011,45(6):59-63.

[13] TIAN Fengping, YANG Ke, CHEN Langnan. Realized volatility forecasting of agricultural commodity futures using the HAR model with time-varying sparsity[J]. International Journal of Forecasting, 2017, 33(1): 132-152.

[14] 张宇翔,赵建民,朱信忠,等. 基于 HDFS 的海量指纹数据云存储优化研究[J]. 浙江师范大学学报:自然科学版,2015,38(2):179-184.

[15] 张 海,马建红. 基于 HDFS 的小文件存储与读取优化策略[J]. 计算机系统应用,2014,23(5):167-171.

[16] HE Hui, DU Zhonghui, ZHANG Weizhe, et al. Optimization strategy of Hadoop small file storage for big data in healthcare[J]. Journal of Supercomputing, 2016, 72(10): 3696-3707.